Non-Representative Surveys: Fast, Cheap, and Mostly Accurate*

Sharad Goel
Stanford University

Adam Obeng
Columbia University

David Rothschild
Microsoft Research

Abstract

Probability-based sampling methods, such as random-digit dialing (RDD) of phones, are a staple of modern survey research and have been successfully used to gauge public opinion for sixty years. Though historically effective, this class of traditional survey techniques are often slow and expensive. At the same time, it has become increasingly quick and cost-effective to collect non-probability-based convenience samples, such as opt-in samples online. Here we investigate the potential of such non-representative data for survey research by administering an online, fully opt-in poll of social and political attitudes. Our survey consisted of 49 multiple-choice attitudinal questions drawn from the probability-based, in-person 2012 General Social Survey (GSS) and select RDD phone surveys by the Pew Research Center. To correct for the inherent biases of non-representative data, we statistically adjust estimates via model-based poststratification. Compared to typical RDD phone polls, the opt-in online survey required less than one-tenth the time and money to conduct. After statistical correction, we find the median absolute difference between the non-probability-based online survey and the probability-based GSS and Pew studies is 7.4 percentage points. Though this difference is considerably larger than if the surveys were all perfect simple random samples, we find the gap is comparable to that between the GSS and Pew estimates themselves, ostensibly because even the best available surveys suffer from substantial non-sampling error. We conclude that non-representative surveys are a promising tool for fast, cheap, and (mostly) accurate measurement of public opinion.

Keywords: Model-based poststratification; non-probability sampling; total survey error

*We thank Andrew Gelman and Matthew Salganik for their helpful comments, and audiences at Columbia, Microsoft Research, and Stanford for their feedback. We also thank Pollfish for use of their survey technology.

Word count: 5,144

1 Introduction

Modern opinion polling is based on the simple and theoretically appealing idea of probability sampling: if each member of the target population has a known, non-zero chance of being surveyed, than a small random sample of the population can be used to accurately estimate the distribution of attitudes in the entire population. This elegant methodological approach has guided polling from the early days of in-home interviewing, through random-digit dialing of landline phones, to more recent mixed-mode polling of landlines and cellphones, and even some online sample. Of course, it has never been possible to reach everyone in the population (e.g., those without permanent addresses), or to guarantee that everyone in the sample responds. Thus, in practice, it is common to use probability-*based* sampling, in which one starts from approximately representative data and then applies a variety of post-sampling adjustments, such as raking [Battaglia et al., 2013], to improve estimates.

The general acceptance of probability-based sampling as the lone legitimate sampling method has permeated the survey research community for over sixty years, and can be traced to a pivotal polling mishap in the 1936 U.S. presidential election campaign. In that race, the popular magazine *Literary Digest* conducted a mail-in survey that attracted over two million responses, a huge sample even by modern standards. The magazine, however, incorrectly predicted a landslide victory for Republican candidate Alf Landon over the incumbent Franklin Roosevelt. Roosevelt, in fact, decisively won the election, carrying every state except for Maine and Vermont. As pollsters and academics have since pointed out, the magazine's pool of respondents was highly biased—consisting mostly of auto and telephone owners, as well as the magazine's own subscribers—and underrepresented Roosevelt's core constituencies [Squire, 1988]. During that same campaign, pioneering pollsters, including George Gallup, Archibald Crossley, and Elmo Roper, used considerably smaller but approximately representative samples to predict the election outcome with reasonable accuracy [Gosnell, 1937]. After Gallup botched the 1948 election between Harry Truman and Thomas Dewey, this early quota sampling would morph into probability-based sampling by 1956; while it had roots in earlier survey research, this is when it became the lone dominant data collection method in public opinion surveys. Accordingly, *non-representative* or *convenience* sampling—catchall phrases that include a variety of non-probability-based data collection strategies—rapidly fell out of favor with polling experts.

Here we revisit the case against non-representative sampling, spurred in part by three recent trends: increased bias in probability-based samples, increased cost in probability-based samples, and decreased cost in non-probability-based samples.

First, there is growing awareness that even the highest-quality representative surveys suffer from a variety of non-sampling errors, and consequently may not be nearly as accurate as generally believed [Shirani-Mehr et al., 2015]. [Shirani-Mehr et al., 2015] shows, empirically, that the error in public opinion polls is about twice as large as sample error. In particular, the extensive literature on *total survey error* [Biemer, 2010, Groves and Lyberg, 2010] points to the need to consider frame, non-response, measurement, and specification error. Frame error occurs when there is a mismatch between the sampling frame and the target population. For example, for phone-based surveys, people without phones would never be included in any sample. Non-response error occurs when missing values are systematically related to the response. For example, as has been recently documented, supporters of a trailing political candidate may be less likely to respond to election surveys [Gelman et al., 2015a]. Measurement error occurs when the survey instrument itself affects the response, often due to order effects [McFarland, 1981] or question wording [Smith, 1987]. Finally, specification error occurs when the concept implied by a survey question differs from what the surveyor seeks to measure. Such errors are particularly problematic when assessing attitudes on social and political issues, which are often hard to pin down precisely. Non-probability-based surveys suffer from the same bias, probably worse [Bethlehem, 2010], but these bias are now a concerning aspect of both probability and non-probability-based samples.

Second, it has become increasingly difficult and expensive to collect representative, or even approximately representative, samples. Random-digit dialing (RDD), the workhorse of modern probability-based polling, has suffered increasingly high non-response rates, in part due to the general public's growing reluctance to answer phone surveys and expanding technical means to screen unsolicited calls [Keeter et al., 2000]. By one study of public opinion surveys, RDD response rates have decreased from 36% in 1997 to 9% in 2012 [Kohut et al., 2012], and other analyses confirm this trend [Council, 2013, Holbrook et al., 2007, Steeh et al., 2001]. Even if the initial pool of targets is representative, those rare individuals who ultimately answer the phone and elect to respond might not be. To combat such issues, the General Social Survey (GSS) employs extreme procedures both to create a comprehensive sampling frame and to reach every subject randomly chosen from the resulting pool. The costs associated with this design, however, are prohibitive for many applications: one iteration of the GSS costs approximately $5 million, about $3 per respondent per question. While there are certainly applications like the GSS

where the added effort is worth the expense, there are certainly applications where it is not.

The third and final trend driving our research is that with recent technological innovations, it is now convenient and cost-effective to collect large numbers of highly non-representative samples via opt-in, online surveys. What took several months for the *Literary Digest* editors to collect in 1936 can now take only a few days with a cost of just pennies per response. And, with graphical interfaces, we break free of the limits and inflexibility of a small postcard, as *Literary Digest* sent, or even the phone which is still the standard mode for probability-based surveys. The challenge, of course, is to extract meaningful signal from these unconventional samples and this is also made easier with increased computing power and increasingly sophisticated software packages.

The literature on this topic has been favorable on the accuracy of probability-based samples versus non-probability-based samples, but the research generally has two crucial differences with this paper. First, most papers, if they examine comparable questions between sample types, examine demographic questions, not opinion questions. This provides a closer approximation of the ground truth for the researchers to compare the results, but respondents answer these types of questions with much more stability than opinion questions. Second, most papers, if they apply analytics to the data, use analytics that are more appropriate for probability-based samples, than non-probability-based samples Yeager et al. [2011]. Further, the literature tends to avoid the questions of cost in both time and money, where there is a wide division between the two sample types.

In this paper, we investigate the speed, cost, and accuracy of non-representative polling by administering and analyzing an online, fully opt-in survey of social and political attitudes. The survey consisted of 14 demographic questions and 49 attitudinal questions that were drawn from the 2012 General Social Survey (GSS) and recent Pew Research Center studies. To correct for the inherent biases of non-representative data, we generate population-level and subgroup-level estimates via model-based poststratification [Gelman and Little, 1997, Wang et al., 2015]. We find that the survey took approximately 2.5 hours to attract 1,000 respondents, and cost approximately \$0.03 per question per respondent. The survey was thus indeed both fast and cheap, requiring less than one-tenth the time and money of traditional RDD polling, and less than one-hundredth the time and money of GSS polling. To gauge accuracy, we compared the statistically corrected poll estimates to those obtained from the GSS and Pew studies. We find the median absolute difference between the non-representative survey and the probability-based GSS/Pew studies is 7.4 percentage points. This difference is considerably larger than expected if all three surveys

were perfect simple random samples. However, perhaps surprisingly, the difference is comparable to that between the GSS and Pew estimates themselves, ostensibly because even these high-quality surveys suffer from substantial total survey error. We conclude that non-representative surveys are a promising tool for quickly, inexpensively, and (mostly) accurately measuring public opinion.

2 Data & Methods

Our primary analysis and results are based on two non-traditional survey methods. First, we conducted an online, non-representative poll on Amazon Mechanical Turk. Second, we conduced a quasi-quota sampling survey administered via mobile phones on the Pollfish survey platform. To gauge the accuracy of these survey methods, we compare our results to those obtained from RDD phone surveys conducted by Pew Research Center, and in-person interviews carried out as part of the 2012 General Social Survey (GSS). We describe our survey collection and analysis methods in more detail below.

2.1 An online, non-representative survey

Amazon Mechanical Turk (AMT) is an online crowd-sourcing marketplace on which individuals and companies can post tasks that workers complete for compensation. AMT was initially used to facilitate the automation of tasks that humans perform well and machines poorly (such as image labeling and audio transcription), but it is increasingly used for social science research [Budak et al., 2015, Flaxman et al., 2015, Paolacci et al., 2010]. We used AMT to conduct a fast, inexpensive, and non-representative survey. Respondents were first asked to answer 14 demographic and behavioral questions (e.g., age, sex, and political ideology), which we primarily used for post-survey adjustment, as described below. Once these were completed, we asked 49 multiple-choice questions on social and public policy (e.g., concerning gay marriage, abortion, and tax policy), in random order, selected from the 2012 GSS and 2012–2014 Pew Research Center RDD phone surveys. As is common practice on AMT [Mason and Suri, 2012, Paolacci et al., 2010], we also asked two "attention questions" (for which there was a clear, correct answer) to confirm that respondents were in fact thoroughly reading and processing the questions; those who failed these checks were not included in the analysis. The full list of survey questions is in the Appendix (see Tables A2 and A3).

The survey was posted on July 6, 2014, and made available to AMT workers who were over 18, resided in the United States, and had a prior record of acceptably com-

pleting more than 80% of tasks attempted. We aimed to recruit 1,000 respondents, a goal that was met in just over 2.5 hours. (For comparison, we note that traditional RDD surveys are typically carried out over several days, and the in-person GSS interviewing process takes three months [Smith et al., 2013, p. vii].) In total, 1,017 respondents started the survey, answering a median number of 46 out of the 49 substantive questions. Respondents were paid $0.05 for every two questions they answered, resulting in a cost per respondent per question approximately 100 times cheaper than the GSS, and approximately 20 times cheaper than traditional RDD polling. The AMT poll was thus indeed both relatively fast and cheap compared to standard probability-based survey methods.

As expected, however, the online, opt-in AMT survey was far from representative. Figure A1 (in the Appendix) shows that respondents deviated significantly from the U.S. population in terms of age, sex, race, education, and political ideology. In particular, relative to the general population, AMT respondents were more likely to be young, male, white, highly-educated, and liberal. These differences likely stem from a variety of inter-related factors, including the need for a computer to use the platform (which results in a wealthier and more educated population of respondents), and heightened interest in our specific task (i.e., a political survey) among certain sub-groups within this population. Regardless of the cause, these discrepancies highlight the need for adjustments to deal with frame and non-response errors that attend surveys in this fully opt-in mode [Couper, 2000].

We employ two popular statistical techniques to correct for the non-representative nature of the AMT survey data: raking and model-based poststratification. Raking [Battaglia et al., 2013] is perhaps the most common approach for adjusting raw survey responses, particularly in probability-based polls. With this method, weights are assigned to each respondent so that the marginal weighted distribution of respondent characteristics match those in the target population. In particular, following DeBell et al. [2010], we assign weights to simultaneously match on five variables: (1) sex; (2) census division; (3) age, categorized as 18–24, 25–30, 30–39, 40–44, 45–49, 50–59, or 60+; (4) race/ethnicity, categorized as white, black, Asian, Hispanic or 'other'; and (5) education, categorized as 'no high school diploma', 'high school graduate', 'some college/associate degree', 'college degree', or 'postgraduate degree'.[1] Marginals in the target population were estimated from the 2012 American Community Survey.

[1]We follow the raking procedure described in DeBell et al. [2010], as implemented in the R package 'anesrake' [Pasek, 2011]. We experimented with several raking procedures, including the method described in Yeager et al. [2011], and found the alternatives yielded comparable, though somewhat worse, performance.

Though popular, raking can suffer from high variance when respondent weights are large, a problem that is particularly acute when the sample is far from representative [Izrael et al., 2009]. Thus, as our primary means of statistical correction, we turn to model-based poststratification (MP) [Gelman and Little, 1997, Ghitza and Gelman, 2013, Park et al., 2004], a technique that has proven effective for correcting non-representative surveys [Wang et al., 2015]. As with raking, MP corrects for known differences between sample and target populations. The idea is to first partition the population into cells (defined by the cross-classification of various attributes of respondents), then use the sample to estimate the mean of a survey variable within each cell, and finally to aggregate the cell-level estimates by weighting each cell by its proportion in the population. In conventional post-stratification, cell values are set to the sample mean. This estimate is unbiased if selection is ignorable (i.e., if sample selection is independent of survey variables conditional upon the variables defining the post-stratification.) The ignorability assumption is more plausible if more variables are conditioned upon. However, adding more variables to the post-stratification increases the number of cells at an exponential rate. If any cell is empty in the sample (which is guaranteed to occur if the number of cells exceeds the sample size), then the conventional post-stratification estimator is not defined. Even for nonempty cells, there can still be problems because sample means are noisy for small cells. Collapsing cells reduces variability, but can leave substantial amounts of selection bias. MP addresses this problem by using regression to obtain stable estimates of cell means.

In our setting, we divide the target population into 53,760 cells based on combinations of sex, age category, race/ethnicity, education, party ID, political ideology, and 2012 presidential vote. For each survey question, we estimate cell means with a multinomial logistic regression model that predicts each individual's response based on the poststratification variables. In particular, the models include seven categorical variables: (1) sex; (2) age, categorized as 18–24, 25–30, 30–39, 40–44, 45–49, 50–59, or 60+; (3) race/ethnicity, categorized as white, black, Asian, Hispanic or 'other'; (4) education, categorized as 'no high school diploma', 'high school graduate', 'some college/associate degree', 'college degree', or 'postgraduate degree'; (5) party ID, categorized as democrat or republican; (6) ideology, categorized as conservative, liberal or moderate; and (7) 2012 presidential vote, categorized as for Obama or Romney. The models additionally include a linear predictor for age so that we can accurately estimate responses for the 60–64 and 65+ age categories, in which we have few respondents. Survey responses are modeled independently for each question (i.e., we fit 49 separate regressions). Given these model-based estimates of cell means, the final poststratification step requires cross-tabulated population data across all of the

variables we consider (so that cell weights can be estimated), for which we turn to the 2012 presidential exit poll. Though exit polls only cover those having voted, they allow us to poststratify based on political variables, which are not recorded in Census Bureau-administered studies like the Current Population Survey or American Community Survey.

2.2 Quasi-quota sampling survey

Though fast and cheap, the fully opt-in survey conducted on AMT was highly non-representative and required extensive statistical correction. As a middle ground between the extreme of AMT and traditional, probability-based polls, we conducted a quasi-quota sampling survey. With quota sampling [Cumming, 1990], respondents are selected so that the sample matches the population on key, pre-specified demographics, such as age and sex. In this case we actively balanced on sex to ensure that sex was representative, but then the polling company randomly sampled from their panel (which may or may not be representative over any other demographic). The survey was conducted on mobile phones via the Pollfish survey platform, a popular tool for conducting such polls. Similar to third-party advertising companies, Pollfish pays mobile application developers to display Pollfish surveys within their applications. To incentivize participation, Pollfish additionally provides bonuses to randomly selected users who complete the surveys.

The survey was launched on December 18, 2014, and was available to individuals over 18 residing in the U.S. who had the Pollfish platform installed on at least one of their mobile phone applications (a population of approximately 10 million people at the time of the study). Given restrictions on survey length, we limited the poll to 12 attitudinal questions (see Table A5 for a full list). We aimed to recruit a gender-balanced pool of 1,000 respondents, and reached this goal in just over 7 hours, with 1,065 respondents completing the full survey of 17 questions (12 attitudinal plus 5 demographic). The retail cost of the survey was $1,500, or $0.08 per respondent per question, about three times as expensive as the AMT survey and about six times cheaper than RDD polling.

2.3 Determining survey accuracy

To evaluate the accuracy of the two survey methods described above, we would ideally like to compare to "ground truth" answers. Finding such a ground truth is difficult, and even enumerative procedures like the U.S. Census have well-known undercoverage bias [Groves and Lyberg, 2010, p. 852], meaning that it is usually

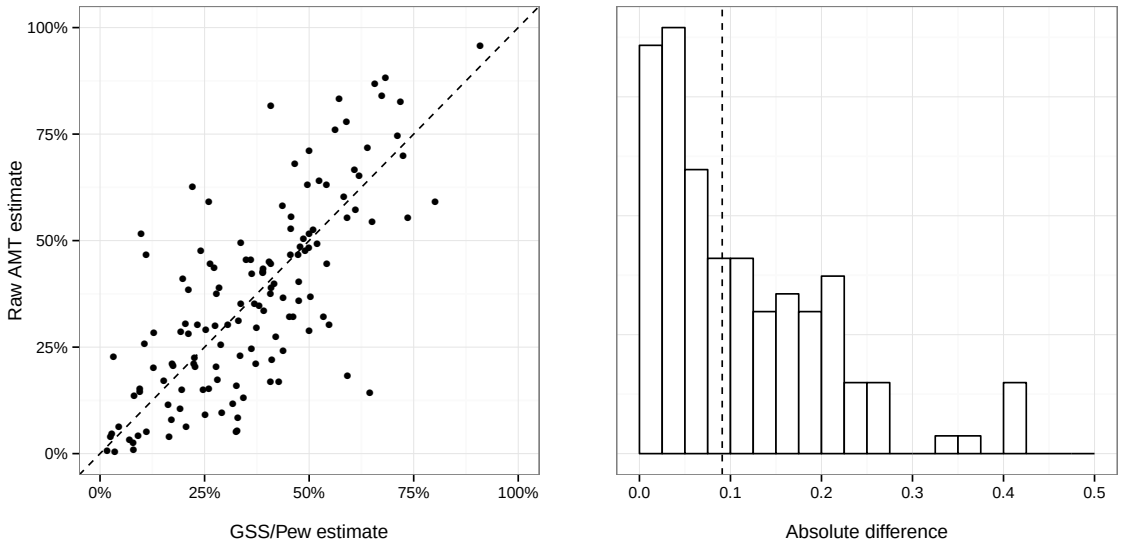
impossible in practice to obtain an error-free estimate of accuracy [Biemer, 2010]. Such difficulties are even more pronounced for the questions of attitude and opinion that interest us here, in part because answers to such questions are rarely, if ever, measured in the full population, and in part because such questions are particularly sensitive to non-sampling errors, such as question order effects [McFarland, 1981]. Moreover, it is often challenging to even identify the underlying construct of interest and design a question to measure that construct [Groves et al., 2013].

Given these issues, we settle for an approximate ground truth as estimated by the GSS and Pew studies, which are regarded to be among the highest quality surveys available. We note that even when ostensibly measuring the same underlying construct (e.g., attitudes on abortion), two different surveys rarely use the exact same wording, an observation that in particular holds for both the GSS and Pew studies. We thus use reasonable judgment to match and compare questions between the surveys. Among the 49 substantive questions we consider, we compare to 13 similar questions asked in the 2012 GSS, and to 36 appearing in a Pew RDD survey conducted in 2012–2014. If a question was asked in multiple Pew studies, we use the most recent survey available. Similarly, in the six cases where a question was asked in both the GSS and by Pew, we compare our estimates to those obtained by Pew, since those surveys were conducted more recently. We further use these six overlapping questions (together with an additional six that appear both on the GSS and Pew surveys, but were not included in ours) to gauge the total survey error of these polls.

3 Results

3.1 Overall accuracy

We start by comparing the raw (i.e., unadjusted) estimates from our online, non-representative survey to estimates obtained from the GSS and Pew, a proxy for the ground truth. Figure 1(a) shows this comparison, where each point in the plot is one of 135 answers to the 49 substantive questions we consider (detailed in Table A2). Figure 1(b) further shows the distribution of differences between the non-representative survey and the approximate ground truth. As indicated by the dashed line, the median absolute difference is 9.1 percentage points, and the RMSE is 15.2. On one hand, this seems like a relatively large gap. On the other hand, given the poll was fully opt-in, conducted on a platform (AMT) with well-known biases, and did not receive the benefit of any statistical adjustment, it is perhaps surprising that the survey was even that accurate.

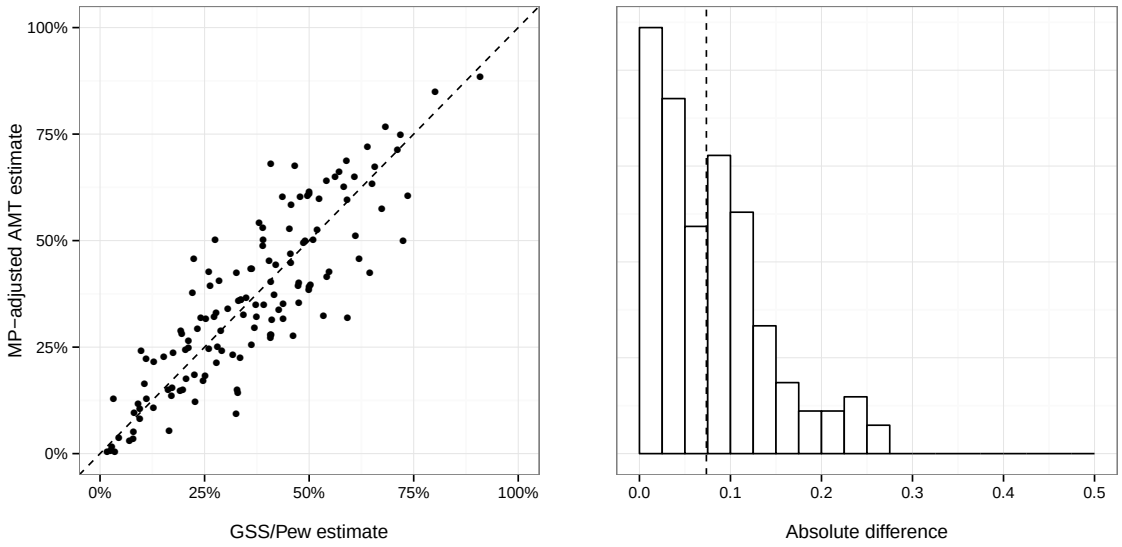


(a) (b)

Figure 1: *Comparison of raw estimates from the online, non-representative poll conducted on Amazon Mechanical Turk to those from the GSS and Pew surveys, a proxy for the ground truth. In panel (a), each point represents an answer (there are 135 answers to 49 questions). In panel (b), the distribution of the differences is shown; the median absolute difference is 9.1 percentage points, indicated by the dashed line.*

Raw survey estimates are a useful starting point for understanding accuracy, but it is just a starting point, even the highest quality surveys—including the GSS and Pew studies—rely on statistical corrections. If we adjust the AMT survey by raking (as described in Section 2.1), we find the median absolute difference between the corrected AMT estimates and the GSS/Pew estimates is 8.7 percentage points, and the RMSE is 13.5. Figure A4 in the Appendix shows the full distribution of differences. The statistical adjustment brings the estimates into somewhat better alignment with one another, though the change is not dramatic.

Finally, Figure 2 compares MP-adjusted estimates from the AMT survey to those from Pew/GSS. After this statistical correction, the median absolute difference between estimates from the non-representative AMT survey and the approximate ground truth is 7.4 percentage points, and the RMSE is 10.2. Notably, the MP-adjusted estimates are more closely aligned with the GSS and Pew studies than the raking-adjusted estimates. As discussed above, this is likely because raking can yield large respondent weights in highly non-representative samples, which in turn



(a) (b)

Figure 2: *Comparison of MP-adjusted estimates from the online, non-representative AMT survey to those from the GSS and Pew surveys. In panel (a), each point represents one of 135 answers to 49 questions. The distribution of the differences between these estimates is shown in panel (b), where the dashed line indicates the median absolute difference of 7.4 percentage points.*

decreases the stability of estimates. Moreover, as can be seen from the distribution of errors in Figure 1(b) and Figure 2(b), the extreme outliers (e.g., those that differ from Pew/GSS by more than 30 percentage points) are no longer present after MP adjustment.

To help put these results into context, we next compare estimates from the GSS to those from the Pew studies on the subset of 12 questions that both ask. As shown in Table 3, the median absolute difference is 8.6 percentage points and the RMSE is 10.1. In particular, the difference between Pew and the GSS is, perhaps surprisingly, comparable to the observed difference (7.4 percentage points) between the AMT survey and these two sources.[2] With appropriate statistical adjustment, the non-

[2]Table 3 shows the difference between MP-adjusted AMT results and the GSS/Pew surveys for the full set of 49 questions. However, we find similar results if we restrict our analysis to the six questions that appear on all three surveys. For example, on this restricted set of questions, the median absolute difference between the MP-adjusted AMT estimates and the Pew studies is 5.8 percentage points, compared to a difference of 5.5 between the GSS and Pew surveys themselves.

	AMT (raw) vs. GSS/Pew	AMT (MP) vs. GSS/Pew	AMT (raking) vs. GSS/Pew	Pollfish vs. Pew	GSS vs. Pew
MAD	9.1	7.4	8.7	7.2	8.6
RMSE	15.2	10.2	13.5	10.6	10.1
# Questions	49	49	49	12	12

Table 3: *Comparison of various data collection and adjustment methodologies. The Pollfish vs. Pew and GSS vs. Pew comparisons are computed over the 12 questions in Table A5; the remaining comparisons are computed over the set of 49 questions in Table A2 . The difference between the MP-adjusted AMT estimates and those from GSS/Pew are on par with the difference between GSS and Pew themselves.*

representative AMT survey aligns about as well with the GSS and Pew surveys as these two high-quality surveys align with one another.

Given that the GSS and Pew surveys are both considered to be among the highest-quality available, why is it that the difference between the two is so large? As discussed in the extensive literature on total survey error [Biemer, 2010, Groves and Lyberg, 2010], there are a variety of non-sampling errors that could explain the discrepancy. First, the surveys are conducted over different modes (in-person for the GSS vs. telephone for the Pew studies). Second, though the GSS and Pew surveys presumably seek to measure the same underlying concepts, the questions themselves are not identically worded. Third, the surveys are not conducted at precisely the same time. Fourth, the GSS uses a fixed ordering of questions, whereas Pew randomizes the order. Fifth, though both the GSS and Pew studies attempt to survey a representative sample of American adults, they undoubtedly reach somewhat different populations, resulting in coverage bias. Sixth, the GSS and Pew likely suffer from different types of non-response, particularly since the surveys are conducted over different modes. Finally, different statistical adjustment procedures are used in each case. Despite these well-known methodological differences, the GSS and Pew surveys are regularly viewed as reasonable approximations of an objective ground truth. That the resulting estimates differ so much highlights the importance of considering non-sampling errors when interpreting survey results.

The fully opt-in AMT poll is arguably at an extreme for non-representative surveys. To investigate the performance of a somewhat more representative, though still non-traditional, data collection methodology, we conducted a quasi-quota sampling survey on the Pollfish mobile phone-based platform. Unlike the GSS and Pew studies, the Pollfish survey is not explicitly attempting to be representative of the U.S. population; however, unlike the AMT survey, some level of representativeness is still

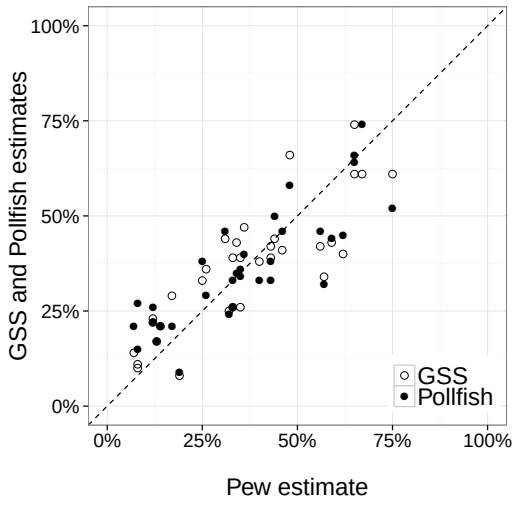


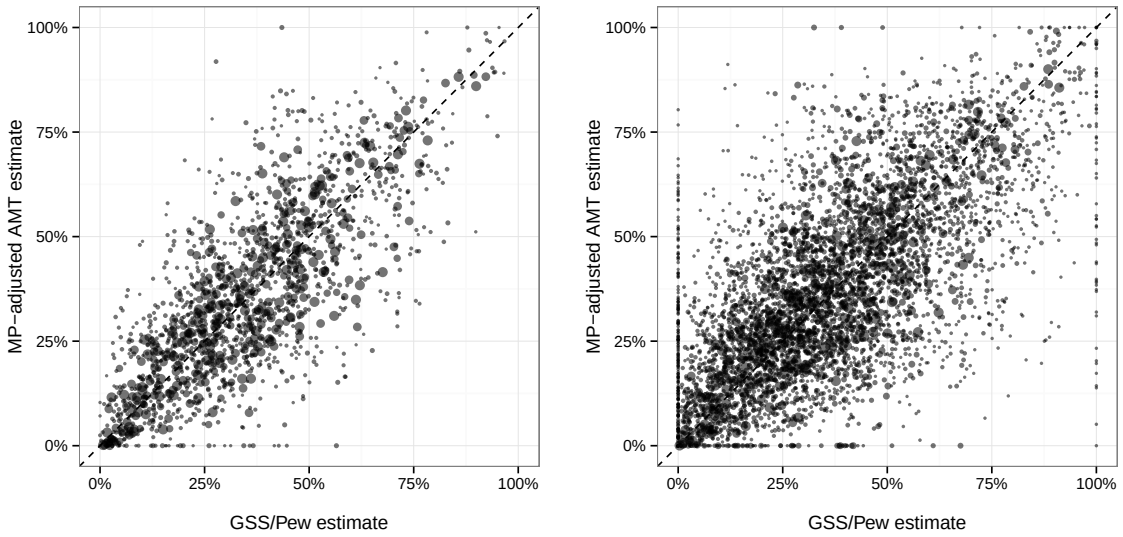
Figure 4: *Comparison of estimates from Pew studies to those from the quasi-quota sampling Pollfish survey (solid circles) and the GSS (open circles). Each point is of one of 33 responses for 12 questions. The Pollfish, GSS, and Pew surveys all yield estimates that are in similar alignment to one another.*

enforced by requiring the pool of respondents to be gender-balanced. We accordingly view Pollfish as a middle ground between the extremes we have thus far considered.

Figure 4 compares results from the GSS, Pew, and Pollfish surveys on the 12 questions that were asked on all three. As is visually apparent from the plot, estimates from the Pollfish survey are about as well-aligned to Pew as are those from the GSS. In quantitative terms, as listed in Table 3, the median absolute difference between the Pollfish and Pew estimates is 7.2 percentage points, whereas the difference between the GSS and Pew is 8.6 percentage points. Thus, we again find that a non-probability-based survey (i.e., Pollfish, in this case) is surprisingly well-aligned with surveys that are generally regarded as the best available.

3.2 Subgroup estimates

We have so far examined overall population-level estimates, finding that after statistical correction non-representative polls are well-aligned with traditional, high-quality surveys. In many cases, however, one not only cares about such top-line results, but also attitudes among various demographic subgroups of the population (e.g., attitudes among liberals, or among 18–24 year-old women). Generating these subgroup estimates is straightforward under both MP-based and raking-based adjustments. In



(a) (b)

Figure 5: *Comparison of subgroup estimates between the MP-adjusted AMT survey and the GSS/Pew studies. In panel (a), each point represents a subgroup based on a single demographic category (e.g., males, or 18–24 year olds). In panel (b), each point represents a subgroup corresponding to a two-way interaction (e.g., male 18–24 year olds, or white women). Points are sized proportional to the size of the subgroup.*

the case of MP, we first use the model to estimate the sample mean in each cell (as before), and then compute a weighted average of the estimates for the cells corresponding to the subgroup of interest. For raking, after assigning the usual weights to each respondent, we simply take a weighted average of respondents in the subgroup.

Figure 5(a) compares MP-adjusted AMT estimates to those from the GSS and Pew for subgroups based on a single demographic category (e.g., males, or 18–24 year olds); Figure 5(b) shows the analogous comparison for subgroups defined by two-way interactions (e.g., 18–24 year-old men, or white women). Subgroup estimates from the AMT, GSS and Pew studies are all likely noisy, but the plots show that they are still generally well-aligned. Specifically, as detailed in Table 6, the median absolute difference between the MP-adjusted AMT estimates and the GSS/Pew studies across all one-dimensional subgroups and the full set of 49 questions is 8.6 percentage points; for comparison, between the GSS and Pew studies themselves (on the 12 questions that both surveys ask) the difference in one-dimensional subgroup estimates is 9.6 percentage points. Similarly for the two-dimensional subgroups, we find

	One-dimensional subgroups			Two-dimensional subgroups		
	AMT (MP) vs. GSS/Pew	AMT (raking) vs. GSS/Pew	GSS vs. Pew	AMT (MP) vs. GSS/Pew	AMT (raking) vs. GSS/Pew	GSS vs. Pew
MAD	8.6	10.5	9.6	10.8	14.2	10.1
RMSE	14.4	17.3	16.9	18.6	24.8	24
# Questions	49	49	12	49	49	12

Table 6: *Comparison of subgroup estimates from the non-representative AMT survey (adjusted with both raking and MP) to those from the GSS and Pew. For both the one- and two-dimensional subgroups, the difference between the MP-adjusted AMT estimates and those from Pew/GSS are on par with the differences between the GSS and Pew studies themselves.*

a difference of 10.8 percentage points for the MP-adjusted AMT estimates versus the GSS/Pew studies, compared to 10.1 for the GSS versus Pew studies.[3] As expected, raking-based estimates are less well-aligned with the GSS and Pew surveys than are the MP-adjusted numbers (see Table 6). Overall, these subgroup-level results are broadly consistent with our top-line analysis in Section 3.1: with appropriate statistical adjustment, non-representative polls yield estimates that differ from high-quality, traditional surveys about as much as these traditional surveys differ from one another.

3.3 The effect of sample size on estimates

We conclude our analysis by looking at how performance of the non-representative AMT survey changes with sample size. To do so, for each sample size k that is a multiple of 50 (between 50 and 1,000), we first randomly sampled k responses from the AMT survey data for each question. On this set of k responses, we then computed MP-adjusted and raking-adjusted estimates. We next compared the adjusted AMT estimates to those form the GSS and Pew surveys (our proxies for the ground truth), computing the median absolute difference. Finally, this entire procedure was

[3]Though the comparison between the AMT and GSS/Pew studies is based on the full set of 49 questions, similar results hold if we restrict to the six questions appearing on all three surveys. In particular, on this smaller set of questions, the median absolute difference between the MP-adjusted AMT estimates and the Pew estimates across all one-dimensional subgroups is 9.6 percentage points, compared to 9.1 for the GSS vs. Pew. Across all two-dimensional subgroups, the analogous numbers are 11.9 for AMT vs. Pew, compared to 12.3 for the GSS vs. Pew.

repeated 20 times to produce expected differences between the adjusted AMT and GSS/Pew estimates for each sample size, with the results plotted in Figure 7. As a baseline for comparison, Figure 7 also shows the difference one would expect if estimates were constructed via (perfect) simple random sampling (SRS).

The plot illustrates three points. First, consistent with our findings above, the MP-based estimates are better aligned to the GSS/Pew results than are raking-based estimates at nearly all sample sizes. This pattern is likely a consequence of high respondent-level raking weights, and accompanying high variance in estimates, that can occur with non-representative samples. Second, even for large sample sizes, the adjusted AMT estimates are not nearly as well-aligned with the GSS and Pew studies as one might expect if these surveys were all conducted with SRS. Third, in contrast to theoretical predictions for SRS, both the MP- and raking-based estimates appear to level-off after a certain sample size, with little apparent change in performance. It is not immediately clear what is ultimately responsible for these latter two phenomena, but we can suggest a possibility. After even a relatively small sample size, bias in the AMT, GSS and Pew estimates (due to, for example, frame and non-response errors) dominate over sampling variation, and thus increasing the number of samples does little to bring the estimates into better alignment.

Discussion

Across a broad range of questions on social and political issues, estimates from our non-representative survey are generally well-aligned with those from the GSS and Pew Research Center studies, the standard-bearers for traditional, probability-based polls. In particular, the difference in estimates from the non-representative and traditional surveys is approximately the same as the difference in estimates between the traditional surveys themselves. This result in part highlights the power of principled, statistical methods to extract signal from non-representative data. However, in at least equal measure, the result also shows that even the best available traditional surveys suffer from substantial total survey error. Our conclusions thus stem both from the surprising accuracy of non-representative surveys, as well as from the surprising inaccuracy of probability-based polls.

Our analysis prompts a natural question: Is it appropriate to interpret the GSS and Pew studies as attempts to measure the same latent quantity? In other words, is the difference between these two a fair benchmark for our results? A savvy decision-maker might attempt to take into account the idiosyncrasies of each survey, including the precise population surveyed, question phrasing, question ordering, survey mode, timing, statistical procedures, and so on. We contend, however, that most end-

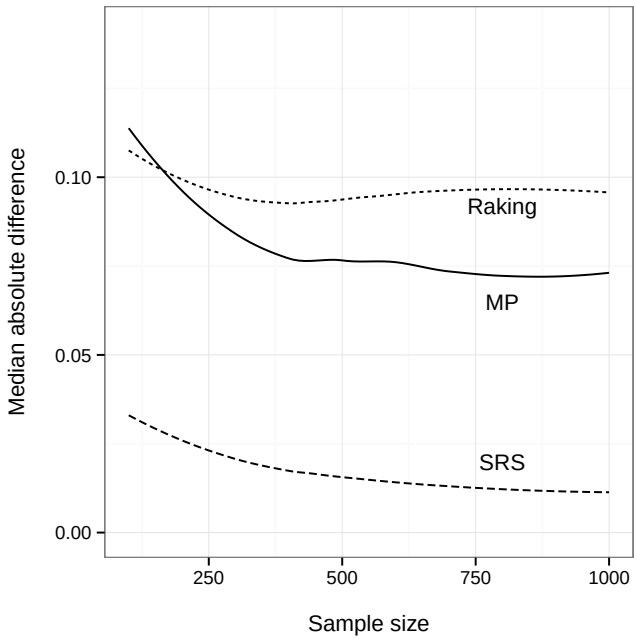


Figure 7: *Median absolute difference between the GSS/Pew studies and the AMT estimates, after correcting the AMT estimate by MP (solid line) and raking (dotted line). For comparison, the dashed line shows the theoretical difference if the estimates were based on perfect simple random samples of the population.*

users are unaware such differences in method exist, and even those who are aware are generally unable to mitigate their effects [Hert, 2003]. As Schuman and Presser [1996, p. 312] note when discussing the effects of question phrasing: "The basic problem is not that every wording change shifts proportions—far from it—but that it is extraordinarily difficult to know in advance which changes will alter marginals and which will not." Given such difficulties, it is not surprising that polls that ostensibly seek to measure the same underlying quantity are often treated as comparable by the media [Frankovic, 2005], despite variance in their procedural details. Thus, at least from the perspective of end-users, it seems appropriate to use the difference in estimates from the GSS and Pew studies as a barometer for our results.

Our non-representative survey consisted exclusively of social and political attitude questions, and so it is unclear how well this approach would work in other domains. At an extreme, it seems difficult—and perhaps impossible—to use an opt-in, online poll to gauge, say, Internet use in the general population, regardless of which statistical methods are applied. A more subtle question is whether non-

representative surveys would be effective in measuring concrete behaviors and traits, which are often less amorphous than attitudes, and which may accordingly be more accurately ascertained by traditional methods. Two recent papers suggest the potential for non-representative surveys in these settings. First, Gelman et al. [2015a] conducted a large, opt-in, non-representative poll on the Xbox gaming platform to track voter intention during the course of the 2012 U.S. presidential election. After statistical correction, they found the poll performed as well, and perhaps even better, than traditional probability-based methods. Second, Yeager et al. [2011] compares probability-based and non-probability polls for estimating "secondary demographics" (e.g., home ownership and household income) and various "non-demographics" (e.g., frequency of smoking and drinking). By comparing to high-quality government statistics, they find the average absolute error of probability-based surveys is 3 percentage points, compared to 5 percentage points for the non-probability methods. The authors of that study conclude that probability-based samples are statistically significantly more accurate than on-probability-based polling, but does that does not mean the difference is meaningful to the end user. Is 2 percentage points of accuracy worth a magnitude or more cost in both time and money? Moreover, we note that the authors adjusted estimates with raking, as is common practice, but more sophisticated model-based poststratification could improve inference from the non-representative data.

With its speed, low-cost, and relative accuracy, non-representative polling offers exciting possibilities for survey research. For example, non-representative surveys can be used to quickly and economically conduct pilot studies for more extensive investigations, which may use a combination of traditional and non-traditional methods. Further, non-representative surveys may facilitate high-frequency, real-time tracking of public opinion [Gelman et al., 2015b]. Though this study is but one data point, our results point to the broad promise of non-representative polls for social research. The savings in time and money is substantial. We do not need to prove the data can be more or even as accurate to make them useful, we suggest a low bar; is the data accurate enough for the end-user or some end-users? Eighty years after the *Literary Digest* failure, non-representative surveys are due for reconsideration, and we hope our work encourages such efforts.

Bibliography

Michael P Battaglia, David C Hoaglin, and Martin R Frankel. Practical considerations in raking survey data. *Survey Practice*, 2(5), 2013.

Jelke Bethlehem. Selection bias in web surveys. *International Statistical Review*, 78 (2):161–188, 2010.

Paul P Biemer. Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74(5):817–848, 2010. ISSN 0033-362X.

Ceren Budak, Sharad Goel, and Justin M Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. Under review, 2015.

National Research Council. *Nonresponse in Social Science Surveys: A Research Agenda*. The National Academies Press, 2013. ISBN 9780309272476. URL `http://www.nap.edu/openbook.php?record_id=18293`.

Mick P Couper. Review: Web surveys: A review of issues and approaches. *Public opinion quarterly*, pages 464–494, 2000. ISSN 0033-362X.

Robert Graham Cumming. Is probability sampling always better? a comparison of results from a quota and a probability sample survey. *Community health studies*, 14(2):132–137, 1990.

Matthew DeBell, Jon A Krosnick, and Arthur Lupia. Methodology report and user's guide for the 2008–2009 anes panel study. Technical report, Stanford University and the University of Michigan, 2010.

Seth R Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. Under review, 2015.

Kathleen A Frankovic. Reporting "the polls" in 2004. *Public Opinion Quarterly*, 69 (5):682–697, 2005. ISSN 0033-362X.

Andrew Gelman and Thomas C. Little. Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 1997. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.44.5270`.

Andrew Gelman, Sharad Goel, Douglas Rivers, and David Rothschild. The mythical swing voter. To appear in the Quarterly Journal of Political Science, 2015a.

Andrew Gelman, Sharad Goel, David Rothschild, and Wei Wang. High-frequency polling with non-representative data. Under review, 2015b.

Yair Ghitza and Andrew Gelman. Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, 57(3):762–776, 2013.

Harold F Gosnell. Technical research how accurate were the polls? *Public Opinion Quarterly*, 1(1):97–105, 1937. ISSN 0033-362X.

Robert M Groves and Lars Lyberg. Total survey error: Past, present, and future. *Public Opinion Quarterly*, 74(5):849–879, 2010. ISSN 0033-362X.

Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. *Survey methodology*. John Wiley & Sons, 2013.

C Hert. Supporting end-users of statistical information: The role of statistical metadata integration in the statistical knowledge network. In *Proceedings of the 2003 National Conference on Digital Government Research*, 2003.

Allyson Holbrook, Jon A Krosnick, Alison Pfent, et al. *The causes and consequences of response rates in surveys by the news media and government contractor survey research firms*, pages 499–528. Wiley, 2007.

David Izrael, Michael P Battaglia, and Martin R Frankel. Extreme survey weight adjustment as a component of sample balancing (aka raking). In *Proceedings from the Thirty-Fourth Annual SAS Users Group International Conference*, 2009.

Scott Keeter, Carolyn Miller, Andrew Kohut, Robert M Groves, and Stanley Presser. Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64(2):125–148, 2000. ISSN 0033-362X.

Andrew Kohut, Scott Keeter, Carroll Doherty, Michael Dimock, and Leah Christian. Assessing the representativeness of public opinion surveys. *Pew Research Center, Washington, DC*, 2012.

Winter Mason and Siddharth Suri. Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23, 2012.

Sam G McFarland. Effects of question order on survey responses. *Public Opinion Quarterly*, 45(2):208–215, 1981.

Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010. ISSN 1930-2975.

David K Park, Andrew Gelman, and Joseph Bafumi. Bayesian multilevel estimation with poststratification: state-level estimates from national polls. *Political Analysis*, 12(4):375–385, 2004.

Josh Pasek. Package 'anesrake'. R Package, 2011.

Howard Schuman and Stanley Presser. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage, 1996.

Houshmand Shirani-Mehr, Sharad Goel, David Rothschild, and Andrew Gelman. Disentangling total error, bias, and variance in election polls. Under review, 2015.

Tom W Smith. That which we call welfare by any other name would smell sweeter an analysis of the impact of question wording on response patterns. *Public Opinion Quarterly*, 51(1):75–83, 1987.

Tom W Smith, Peter Marsden, Michael Hout, and Jibum Kim. *General Social Surveys, 1972-2012*. Chicago: National Opinion Research Center [producer] and Storrs, CT: The Roper Center for Public Opinion Research, University of Connecticut [distributor], 2013.

Peverill Squire. Why the 1936 literary digest poll failed. *Public Opinion Quarterly*, 52(1):125–133, 1988. ISSN 0033-362X.

Charlotte Steeh, Nicole Kirgis, Brian Cannon, and Jeff DeWitt. Are they really as bad as they seem? nonresponse rates at the end of the twentieth century. *Journal of Official Statistics*, 17(2):227–248, 2001.

Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, pages 980–991, 2015.

David S Yeager, Jon A Krosnick, LinChiat Chang, Harold S Javitz, Matthew S Levendusky, Alberto Simpser, and Rui Wang. Comparing the accuracy of rdd telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 2011. ISSN 0033-362X.

Appendix

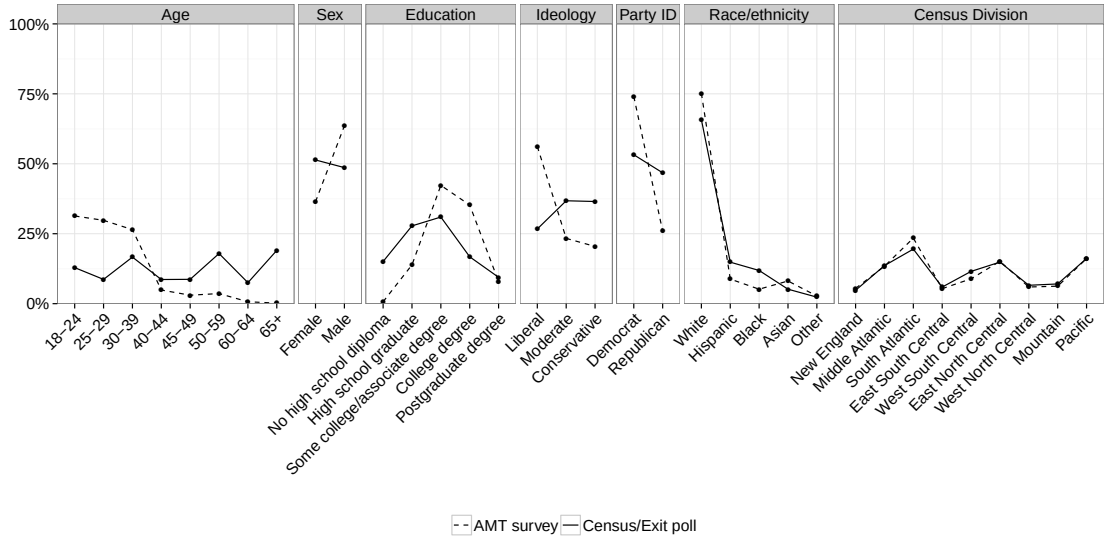


Figure A1: *Comparison of Amazon Mechanical Turk respondent characteristics to those of the general American population, as estimated by the 2012 American Communities Survey (1% sample) and the 2012 presidential exit polls. Relative to the general population, the opt-in AMT survey respondents are younger, more educated, more liberal, and more often male.*

Survey	ID	Topic of Questions
GSS	2012/bible	Belief in Bible
GSS	2012/cappun	Death penalty
GSS	2012/courts	Courts and criminals
GSS	2012/divlaw	Ease of divorce
GSS	2012/happy	Current happiness
GSS	2012/letdie1	Medical assisted suicide
GSS	2012/letin1	Quantity of immigrants
GSS	2012/pillok	Birth control for teenagers
GSS	2012/pornlaw	Pornography laws
GSS	2012/prayer	Separation of church and state
GSS	2012/sexeduc	Sexual education

Continued on next page

Table A2 – continued from previous page

Survey	ID	Topic of Questions
GSS	2012/sprtprsn	Personal spirituality
GSS	2012/tax	Federal income tax rate
Pew	2012.01/early/q17c	Fault of racial disparity
Pew	2012.04/q41df2	Existence of God
Pew	2012.04/q45fb	Trust other people
Pew	2013.02/q28af1	Healthcare spending
Pew	2013.02/q28gf1	Social Security spending
Pew	2013.02/q28hf1	Highway and bridges spending
Pew	2013.02/q28if1	Assistance to poor spending
Pew	2013.03/q15e	Corporate profits
Pew	2013.03/q6f1	Debt reduction versus assistance to elderly
Pew	2013.05/q17a	Difficulty of being poor
Pew	2013.05/q17b	Homosexuality
Pew	2013.05/q17c	Islam and violence
Pew	2013.05/q40	Gun control
Pew	2013.05/q53	Gun ownership
Pew	2013.06/q1	Community as place to live
Pew	2013.06/q25	Economic conditions one year from now
Pew	2013.06/q26	Current personal financial situation
Pew	2013.06/q46	Undocumented immigrants
Pew	2013.07/q1	Current direction of country
Pew	2013.07/q10	Anti-terror policy
Pew	2013.07/q2	Obama job approval
Pew	2013.07/q33	Current economic conditions
Pew	2013.07/q40	Abortion
Pew	2013.07/q50	News organizations and facts
Pew	2013.07/q7a	Republican party favorability
Pew	2013.07/q7b	Democratic party favorability
Pew	2013.07/q7c	Supreme Court favorability
Pew	2013.07/q7d	Congress favorability
Pew	2013.07/q9	Government and terrorism prevention
Pew	2013.07/teaparty2	Tea Party
Pew	2014/polarization/q25a	Government efficiency
Pew	2014/polarization/q25b	Government regulation
Pew	2014/polarization/q25d	Debt reduction versus assistance to poor

Continued on next page

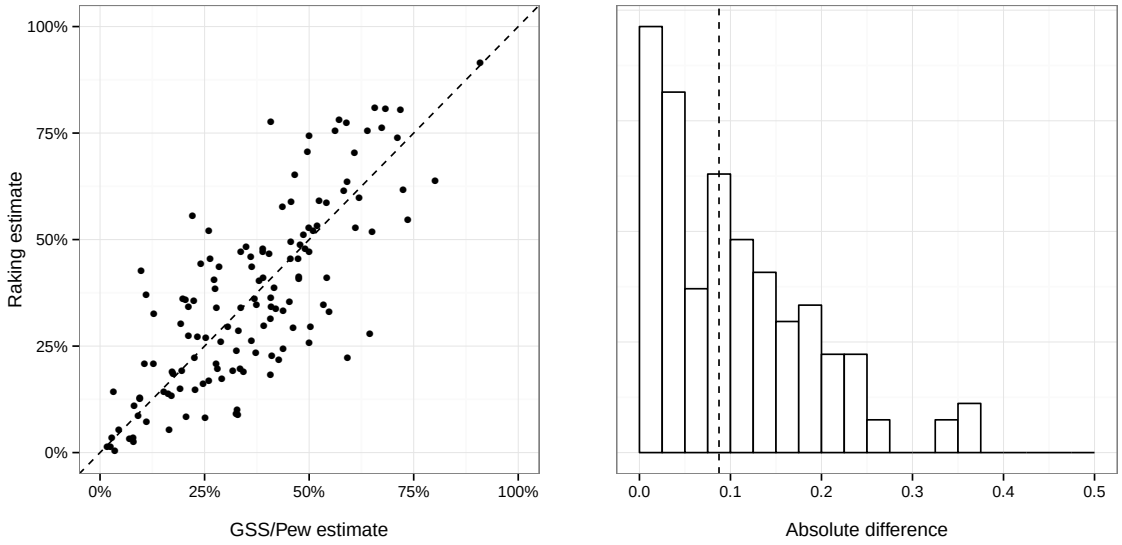
Table A2 – continued from previous page

Survey	ID	Topic of Questions
Pew	2014/polarization/q25g	Value of immigrants
Pew	2014/polarization/q25i	Best way to keep peace
Pew	2014/polarization/q25r	Environmental regulation

Table A2: *List of 49 substantive questions asked in the AMT survey, from GSS and Pew studies.*

Order	Topic of Question
1	Age
2	Gender
3	Race
4	Hispanic or Latino
5	State
6	Zip Code
7	Education completed
8	Education expected
9	Political ideology
10	Party identity
11	Strength of party identity
12	2012 presidential turnout to vote
13	2012 presidential vote
14	Word games

Table A3: *Demographic questions asked of all the AMT survey participants.*



(a) (b)

Figure A4: *Comparison of raking-adjusted estimates from the online, non-representative AMT poll to those from the GSS and Pew surveys. In panel (a), each point represents one of 135 answers to 49 questions. The distribution of the differences between these estimates is shown in panel (b), where the dashed line indicates the median absolute difference of 8.7 percentage points.*

Topic:Response	GSS	Pew	Difference
Military: Keep spending the same	0.43	0.42	0.01
Military: Decrease spending	0.32	0.25	0.08
Military: Increase spending	0.25	0.33	0.09
Crime: Keep spending the same	0.34	0.43	0.09
Crime: Decrease spending	0.07	0.14	0.07
Crime: Increase spending	0.59	0.43	0.16
Education: Keep spending the same	0.17	0.29	0.12
Education: Decrease spending	0.08	0.10	0.02
Education: Increase spending	0.75	0.61	0.14
Environment: Keep spending the same	0.31	0.44	0.13
Environment: Decrease spending	0.12	0.22	0.10
Environment: Increase spending	0.57	0.34	0.23
Foreign Aid: Keep spending the same	0.26	0.29	0.03
Foreign Aid: Decrease spending	0.66	0.49	0.17
Foreign Aid: Increase spending	0.08	0.22	0.14
Health: Keep spending the same	0.26	0.36	0.10
Health: Decrease spending	0.12	0.23	0.11
Health: Increase spending	0.62	0.40	0.21
Highway and bridges: Keep spending the same	0.44	0.44	0.00
Highway and bridges: Decrease spending	0.13	0.17	0.04
Highway and bridges: Increase spending	0.43	0.39	0.04
Science: Keep spending the same	0.46	0.41	0.05
Science: Decrease spending	0.14	0.21	0.07
Science: Increase spending	0.40	0.38	0.02
Social Security: Keep spending the same	0.36	0.47	0.12
Social Security: Decrease spending	0.08	0.11	0.02
Social Security: Increase spending	0.56	0.42	0.14
Gun owner: no	0.65	0.61	0.04
Gun owner: yes	0.35	0.39	0.04
Fault for racial disparity: Blacks who can't get ahead in this country	0.65	0.74	0.09
Fault for racial disparity: Racial discrimination is the main reason	0.35	0.26	0.09
Trust other people: Can't be too careful	0.67	0.61	0.06
Trust other people: Most people can be trusted	0.33	0.39	0.06

Table A5: *Results from the GSS and Pew for the 12 comparable questions asked in both. For the nine budget-related questions, the Pew studies used the following format: "If you were making up the budget for the federal government this year, would you increase spending, decrease spending or keep spending the same for [question topic]." For the GSS, the format for these nine questions was: "We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount. First [question topic], are we spending too much, too little, or about the right amount on [question topic]."*