Online, Opt-in Surveys: Fast and Cheap, but are they Accurate?

Sharad Goel
Stanford University
scgoel@stanford.edu

Adam Obeng
Columbia University
adam.obeng@columbia.edu

David Rothschild
Microsoft Research
davidmr@microsoft.com

ABSTRACT

It is increasingly common for government and industry organizations to conduct online, opt-in surveys, in part because they are typically fast, inexpensive, and convenient. Online polls, however, attract a non-representative set of respondents, and so it is unclear whether results from such surveys generalize to the broader population. These non-representative surveys stand in contrast to probability-based sampling methods, such as random-digit dialing (RDD) of phones, which are a staple of traditional survey research. Here we investigate the accuracy of non-representative data 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, a classic statistical tool but one that is only infrequently used for bias correction. Our online survey took less than one-twentieth the time and money of traditional RDD polling, and less than one-hundredth the time and money of GSS polling. 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 percentage points. This difference is considerably larger than if the surveys were all perfect simple random samples drawn from the same population; the gap, however, is comparable to that between the GSS and Pew estimates themselves. Our results suggest that with proper statistical adjustment, online, non-representative surveys are a valuable tool for practitioners in varied domains.

1 INTRODUCTION

Traditional 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 to some online surveys. 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 [23, 35], to improve estimates.

Within the survey research community, adoption of probability-based methods can be traced to a pivotal polling mishap in the 1936 U.S. presidential election campaign.[1] 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 [33]. During that same campaign, pioneering pollsters, including George Gallup, Archibald Crossley, and Elmo Roper, used considerably smaller but approximately representative quota samples to predict the election outcome with reasonable accuracy [15]. By 1956, quota sampling matured into our contempoary notion of probability-based sampling, and alternative *non-representative* or *convenience* sampling methods—catchall phrases that include a variety of non-probability-based data collection strategies—fell out of favor with polling experts.

The last sixty years has seen significant advances in both data collection and statistical methodology, prompting us to revisit the case against non-representative sampling methods. 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 [13, 36]. Model-based poststratification is a classic statistical method for reducing variance, but its use for bias correction is relatively new.

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-twentieth 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 and Pew studies is 7 percentage points. This difference is considerably larger than expected if

Working paper, Stanford University
2017. 978-x-xxxx-xxxx-x/YY/MM.
DOI: 10.1145/nnnnnnn.nnnnnnn

[1] The idea of probability sampling predates its use in the 1936 election (c.f. Bowley [3]), but this election was an important success in the history of such methods.

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* [17]. These results suggest that non-representative surveys can be valuable to quickly and inexpensively measure attitudes with a degree of accuracy that may be acceptable for many applications, including those in public policy, marketing, and beyond.

2 RELATED WORK

Our work is spurred by three recent trends: (1) growing awareness that probability-based surveys suffer from large, and possibly increasing, non-sampling errors; (2) increasing cost of probability-based surveys; and (3) decreasing cost of non-probability-based sampling. We discuss each of these in turn below.

First, the extensive literature on *total survey error* [2, 17] points to the need to consider errors that arise from sources other than sampling variation. It is now well known that even the highest quality probability-based surveys suffer from these non-sampling errors, and consequently may not be nearly as accurate as generally believed. For example, Shirani-Mehr et al. [30] show that the empirical error in election polls is about twice as large as theoretical estimates based only on sampling variation. Such work specifically notes the importance of frame, non-response, measurement, and specification errors. 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 [11]. Measurement error occurs when the survey instrument itself affects the response, often due to order effects [25] or question wording [31]. 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 opinions and attitudes, which are often hard to pin down precisely. We note that non-probability-based surveys suffer from these same biases, probably even more so [1], but it is now understood that such issues are not limited to convenience 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, suffers from 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 [21]. By one study of public opinion surveys, RDD response rates have decreased from 36% in 1997 to 9% in 2012 [22], and other analyses confirm this trend [5, 19, 34]. Even if the initial pool of targets is representative, those individuals who ultimately answer the phone and elect to respond might not be. To combat such issues, the General Social Survey (GSS) employs elaborate 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. Although there are certainly applications like the GSS where the added effort is worth the expense, there are also many 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, online polls can expand upon the types of questions that can be asked on a small postcard, as *Literary Digest* sent, or asked over 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. As we describe below, this task is made easier with advances in statistical theory.

To help position our paper in the ongoing academic discussion about non-probability-based survey methods, we briefly highlight three key differences between our approach and that of past work. First, in comparing the accuracy of probability-based and non-probability-based surveys, past studies have primarily examined demographic and behavioral questions (e.g., smoking frequency) rather than attitudinal questions (e.g., views on a product) [37]. Demographic and behavioral questions have the advantage that their answers are often known with high accuracy, for example through a government census; however, such questions are less susceptible to non-sampling errors that often afflict the type of attitude questions that are central to many investigations. Second, when correcting non-representative samples, past work has generally applied statistical methods designed for probability-based samples—such as raking—rather than techniques tailored to the specific challenges of convenience samples. Finally, the literature has largely avoided examing the inherent tradeoff between survey accuracy and cost, both in terms of time and money. By addressing these considerations, in this paper we seek to more fully evaluate the potential of non-representative sampling for practioners.

3 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 conducted 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.

3.1 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 [4, 9, 26]. 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 [24, 26], 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 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 completing 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 [32, 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 both fast and cheap compared to standard probability-based survey methods.

As expected, however, the online, opt-in AMT survey was far from representative, with respondents deviating 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 interrelated 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 subgroups 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 [6].

3.2 Statistical adjustment

We employ two techniques to statistically correct for the non-representative nature of the AMT survey data: raking and model-based poststratification.

Raking [23] 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 weighted distribution of respondent characteristics match those in the target population. For a sample of n individuals $x_1, \ldots, x_n$, we denote by $x_{ij} \in \{0, 1\}$ the j-th trait of the i-th individual. For example, x_{ij} may equal 1 if the individual is female and zero otherwise; categorical traits, such as race, are encoded as a series of binary indicator variables. Given a set of target values c_j that specify the prevalence of each trait in the target population, raking attempts to find respondent weights w_i so that

$$c_j = \frac{\sum_{i=1}^{n} w_i x_{ij}}{\sum_{i=1}^{n} w_i} \qquad \forall j.$$

Survey responses y_i are then accordingly weighted to yield the raking estimate:

$$\hat{y}^{\text{rake}} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}.$$

Following DeBell et al. [8], 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". Target values c_j were estimated from the 2012 American Community Survey. To carry out this procedure, we used the R package `anesrake` [28].[2]

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 [20]. Thus, as our primary means of statistical correction, we turn to *model-based poststratification* (MP) [13, 14, 27], a technique that has recently proven effective for correcting non-representative surveys [36]. As with raking, MP corrects for known differences between sample and target populations. The core idea is to partition the population into cells (e.g., based on combinations of various demographic attributes), use the sample to estimate the response variable within each cell, and finally to aggregate the cell-level estimates up to a population-level estimate by weighting each cell by its relative proportion in the population.

The poststratification estimate is defined by,

$$\hat{y}^{\text{MP}} = \frac{\sum_{j=1}^{J} N_j \hat{y}_j}{\sum_{j=1}^{J} N_j} \tag{1}$$

where $\hat{y}_j$ is the estimate of y in cell j, and N_j is the size of the j-th cell in the population. We can analogously derive an estimate of y at any subpopulation level s (e.g., attitudes among young men) by

$$\hat{y}_s^{\text{MP}} = \frac{\sum_{j \in J_s} N_j \hat{y}_j}{\sum_{j \in J_s} N_j} \tag{2}$$

where J_s is the set of all cells that comprise s. As is readily apparent from the form of the poststratification estimator, the key is to obtain accurate cell-level estimates, as well as estimates for the cell sizes.

One popular way to generate cell-level estimates is to simply average sample responses within each cell. If we assume that within a cell the sample is drawn at random from the larger population, this yields an unbiased estimate. However, this assumption of cell-level simple random sampling is only reasonable when the partition is sufficiently fine; on the other hand, as the partition becomes finer, the cells become sparse, and the empirical sample averages become unstable. We address these issues by instead generating cell-level estimates via a regression model that predicts survey response conditional on demographic attributes.

[2] We experimented with several raking procedures, including the method described in Yeager et al. [37], and found the alternatives yielded comparable, though somewhat worse, performance.

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, 4–49, 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+ age category, 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 responses within each cell, 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 the American Community Survey.

To facilitate use of model-based poststratification by practitioners, we are releasing our R source code to implement this procedure. We are also developing an R package, `postr`, to further ease adoption of the method.[3]

3.3 Quasi-quota sampling survey

Though fast and cheap, the fully opt-in survey we 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 on mobile phones via the Pollfish survey platform, a popular tool for administering such polls. With quota sampling [7], 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, but otherwise randomly sampled individuals from the Pollfish panel. 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.

Our 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. 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.

[3]Our code takes as input any user-specified postratification weights, such as those derived from the U.S. Census. The weights we use here come from proprietary 2012 exit poll data, and so cannot be re-distributed.

3.4 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 [17, p. 852], meaning that it is usually impossible in practice to obtain an error-free measure of accuracy [2]. 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 [25]. Moreover, it is often challenging to even identify the underlying construct of interest and design a question to measure that construct [16].

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 climate change), 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.

4 RESULTS

4.1 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 (top panel) shows this comparison, where each point in the plot is one of 135 answers to the 49 substantive questions we consider. Figure 1 (bottom panel) 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. As expected, this is a relatively large gap; however, given the poll was fully opt-in, was 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.

Raw survey estimates are a useful starting point for understanding accuracy, but even the highest quality surveys—including the GSS and Pew studies—rely on statistical corrections. If we adjust

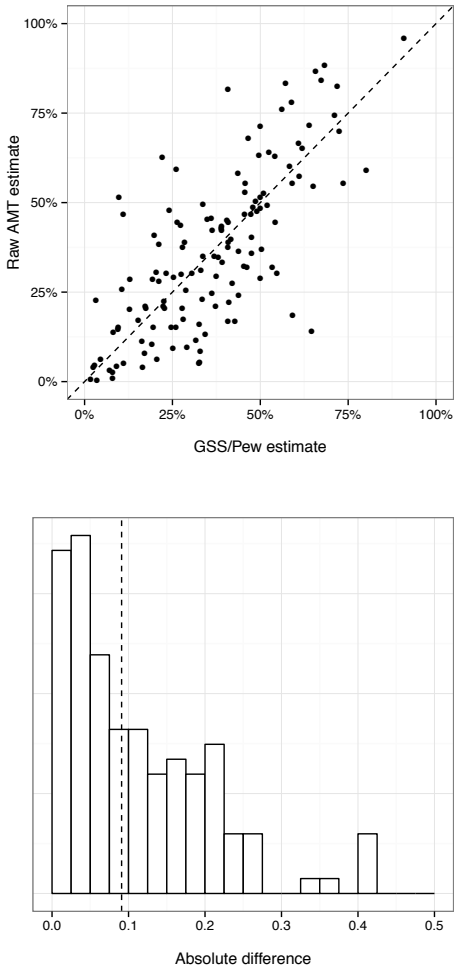

Figure 1: Comparison of raw estimates from the online, non-representative poll conducted on Amazon Mechanical Turk to those from the GSS and Pew surveys, a proxy for the ground truth. Top: each point represents an answer (there are 135 answers to 49 questions). Bottom: the distribution of the differences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

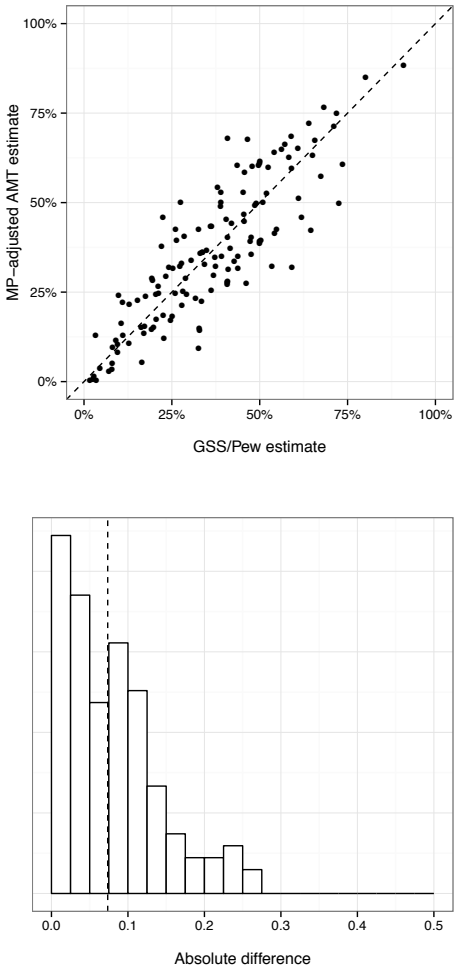

Figure 2: Comparison of MP-adjusted estimates from the online, non-representative AMT survey to those from the GSS and Pew surveys. In the top panel each point represents one of 135 answers to 49 questions. The distribution of the differences between these estimates is shown in the bottom panel, where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

the AMT survey by raking (as described in Section 3.1), 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. 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. 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 decreases the stability of estimates. Moreover, as can be seen from the distribution of errors in the bottom panels of Figure 1 and Figure 2, 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 1, 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

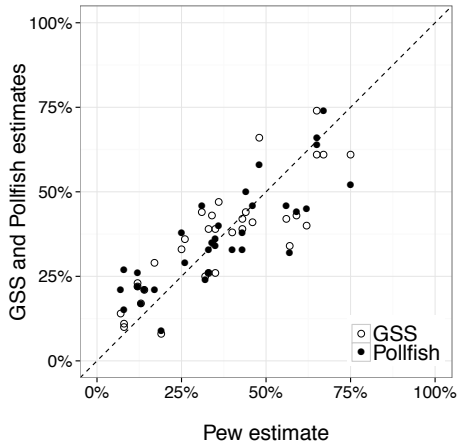

Figure 3: Comparison of estimates from Pew studies to those from the quasi-quota sampling Pollfish survey (solid circles) and the GSS (open circles). Each point is of one of 33 responses for 12 questions. The Pollfish, GSS, and Pew surveys all yield estimates that are in similar alignment to one another.

surprisingly, comparable to the observed difference (7.4 percentage points) between the AMT survey and these two sources.[4] With appropriate statistical adjustment, the non-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 [2, 17], 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.

[4]Table 1 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.

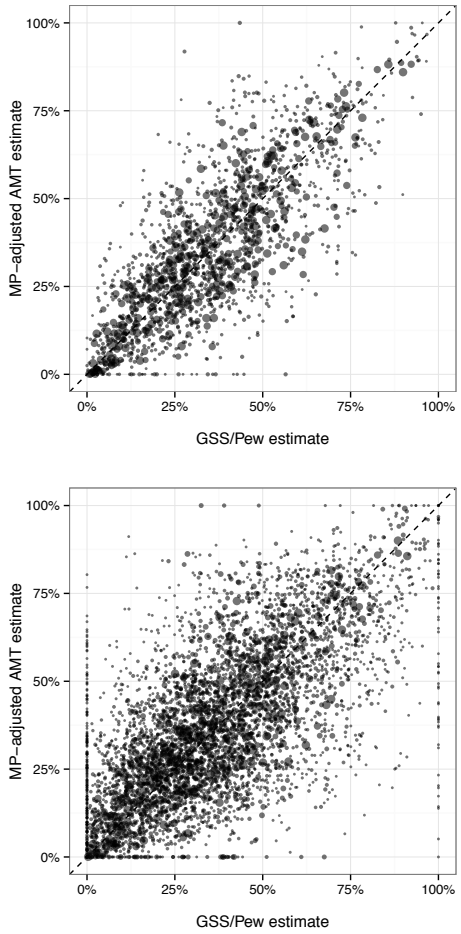

Figure 4: Comparison of subgroup estimates between the MP-adjusted AMT survey and the GSS/Pew studies. Top: each point represents a subgroup based on a single demographic category (e.g., males, or 18–24 year olds). Bottom: each point represents a subgroup corresponding to a two-way interaction (e.g., male 18–24 year olds, or white women). Points are sized proportional to the size of the subgroup.

The 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 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 3 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

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

as well-aligned to Pew as are those from the GSS. In quantitative terms, as listed in Table 1, 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 among the best available.

4.2 Subgroup estimates

We have so far examined overall population-level estimates, finding that after statistical correction non-representative polls are reasonably well-aligned with traditional, high-quality surveys. In many cases, however, one not only cares about such top-line results, but also about 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 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; Eq. (2) makes this precise. For raking, after assigning the usual weights to each respondent, we take a weighted average of respondents in the subgroup.

Figure 4 (top panel) 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 4 (bottom panel) 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 2, 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 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.[5] As before, raking-based estimates are less well-aligned with the GSS and Pew surveys than are the MP-adjusted numbers (see Table 2). Overall, these subgroup-level results are broadly consistent with our top-line analysis in Section 4.1: 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.

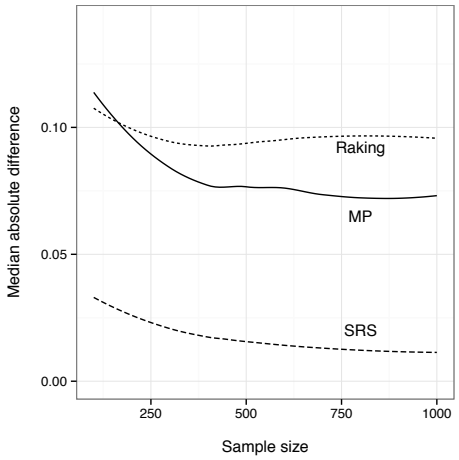

Figure 5: Median absolute difference between the GSS/Pew studies and the AMT estimates, after correcting the AMT estimate by MP (solid line) and raking (dotted line). For comparison, the dashed line shows the theoretical difference if the estimates were based on perfect simple random samples of the population.

4.3 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

[5] 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.

	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 2: 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.

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 from the GSS and Pew surveys, computing the median absolute difference. Finally, this entire procedure was 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 5. As a baseline for comparison, Figure 5 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 attitude and opinion questions, we find that the difference in estimates between the non-representative and traditional surveys we examine is approximately the same as the difference in estimates between the traditional surveys themselves. This result in part highlights the value of principled, statistical methods to extract signal from non-representative data. In at least equal measure, the result also shows that even the best available traditional surveys suffer from substantial total survey error.

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-users are unaware such differences in method exist, and even those who are aware are generally unable to mitigate their effects [18]. As Schuman and Presser [29, 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 [10], 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. For example, Yeager et al. [37] 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-based methods.[6]

With its speed, low-cost, and relative accuracy, online opt-in polling offers promise for survey research in a variety of applied settings. 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

[6]The authors adjusted estimates with raking, as is common practice, but model-based poststratification might have improved estimates from the non-representative data.

facilitate high-frequency, real-time tracking of sentiment [12]. To be clear, non-representative polls should not be viewed as a replacement for traditional survey methods. Many important applications require the highest quality estimates, justifying the added expense of probability-based methods. However, our findings point to the potential of non-representative polls to complement traditional approaches to social research. Eighty years after the *Literary Digest* failure, non-representative surveys are due for reconsideration, and we hope our work encourages such efforts.

ACKNOWLEDGMENTS

We thank Andrew Gelman for helpful conversations.

REFERENCES

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

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

[3] Arthur Lyon Bowley. 1906. Address to the economic science and statistics section of the british association for the advancement of science, york, 1906. *Journal of the Royal Statistical Society* 69, 3 (1906), 540–558.

[4] Ceren Budak, Sharad Goel, and Justin M Rao. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly* (2016).

[5] National Research Council. 2013. *Nonresponse in Social Science Surveys: A Research Agenda.* The National Academies Press. http://www.nap.edu/openbook.php?record_id=18293

[6] Mick P Couper. 2000. Review: Web surveys: A review of issues and approaches. *Public opinion quarterly* (2000), 464–494.

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

[8] Matthew DeBell, Jon A Krosnick, and Arthur Lupia. 2010. *Methodology report and user's guide for the 2008–2009 ANES Panel Study.* Technical Report. Stanford University and the University of Michigan.

[9] Seth R Flaxman, Sharad Goel, and Justin M Rao. 2016. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly* (2016).

[10] Kathleen A Frankovic. 2005. Reporting "the polls" in 2004. *Public Opinion Quarterly* 69, 5 (2005), 682–697.

[11] Andrew Gelman, Sharad Goel, Douglas Rivers, and David Rothschild. 2016. The Mythical Swing Voter. *Quarterly Journal of Political Science* (2016).

[12] Andrew Gelman, Sharad Goel, David Rothschild, and Wei Wang. 2016. High-Frequency Polling with Non-Representative Data. (2016). Under review.

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

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

[15] Harold F Gosnell. 1937. TECHNICAL RESEARCH HOW ACCURATE WERE THE POLLS? *Public Opinion Quarterly* 1, 1 (1937), 97–105.

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

[17] Robert M Groves and Lars Lyberg. 2010. Total survey error: Past, present, and future. *Public Opinion Quarterly* 74, 5 (2010), 849–879.

[18] C Hert. 2003. 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.*

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

[20] David Izrael, Michael P Battaglia, and Martin R Frankel. 2009. Extreme Survey Weight Adjustment as a Component of Sample Balancing (aka Raking). In *Proceedings from the Thirty-Fourth Annual SAS Users Group International Conference.*

[21] Scott Keeter, Carolyn Miller, Andrew Kohut, Robert M Groves, and Stanley Presser. 2000. Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly* 64, 2 (2000), 125–148.

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

[23] Sharon Lohr. 2009. *Sampling: Design and Analysis.* Nelson Education.

[24] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23.

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

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

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

[28] Josh Pasek. 2011. Package 'anesrake'. R Package. (2011).

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

[30] Houshmand Shirani-Mehr, Sharad Goel, David Rothschild, and Andrew Gelman. 2017. Disentangling Total Error, Bias, and Variance in Election Polls. (2017). Under review.

[31] Tom W Smith. 1987. 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 (1987), 75–83.

[32] Tom W Smith, Peter Marsden, Michael Hout, and Jibum Kim. 2013. *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].

[33] Peverill Squire. 1988. Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly* 52, 1 (1988), 125–133.

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

[35] D Stephen Voss, Andrew Gelman, and Gary King. 1995. A review: preelection survey methodology: details from eight polling organizations, 1988 and 1992. *Public Opinion Quarterly* (1995), 98–132.

[36] Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. Forecasting Elections with Non-Representative Polls. *International Journal of Forecasting* (2015), 980–991.

[37] David S Yeager, Jon A Krosnick, LinChiat Chang, Harold S Javitz, Matthew S Levendusky, Alberto Simpser, and Rui Wang. 2011. Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly* (2011).