Validating the COVID States Method: A comparison of non-probability and probability-based survey methods

Introduction

Probability-based sampling is considered the gold standard in public opinion polls because, by giving everyone in the population a known and equal chance to participate, it provides a theoretical rather than observational means for calculating the margin of error. In practice, probability-based samples are also more likely to be representative of a population than more convenient samples. However, as the 2016 and 2020 American presidential elections have shown, well-grounded precision is not a cure-all for methodological soundness. If basic methodological concerns like selection bias, mode effects, or question framing go unaddressed and model weights are used inappropriately, this theoretically sound precision only works to give us estimates that are only precisely incorrect.

Recruiting participants to take a survey using probability-based sampling has never been more costly and difficult as response rates have fallen below 10% while the prevalence of cell phones has made geographic sampling by area code more difficult. Moreover, to the extent that non-response bias is driven by core social factors like age, race, political ideology, education, income, gender, and the like; it represents a significant threat to the generalizability of results from surveys conducted using probability-based sampling. We are working harder and harder for a method that provides much less certainty than we think.

In contrast, recruiting participants using non-probability-based methods has never been easier. Sample aggregators, crowdwork platforms, and online advertising make it easy to recruit people online quickly. This scalability and cost-effectiveness are traded off against two basic problems. First, these samples are much more likely to be impacted by selection bias in who receives recruitment requests and who decides to participate. Second, the online samples tend to produce lower-quality data as people speed through or potentially game the survey for pay. The question is whether our ability to recruit more subjects more quickly and cheaply can compensate for the selection bias and lower data quality.

In this study, we compare biases and data quality in three methods of non-probability, online subject recruitment – crowd workers from Mechanical Turk (MTurk), volunteers from Facebook Ads, and online samples from PureSpectrum. This recruitment is from a larger project of monthly surveys beginning from April 2020. This paper will focus on parallel recruitment efforts from September 9, 2020 – September 29, 2020.

Recruitment procedures

PureSpectrum

We recruited participants from PureSpectrum, a sample reseller allowing researchers to recruit large samples of study participants based on a wide variety of specific demographics. PureSpectrum has been recruiting roughly 20,000 participants every month meeting state-level census-based quotas for age, gender, and race to produce national and state-level estimates of key behavioral, attitudinal, and health statistics during the Covid 19 pandemic for the Covid States project. We recruited from PureSpectrum in September and the interquartile range for time to completion in September was 15 to 28 minutes with a median of 21.

Mechanical Turk

In the September wave, we posted HITs on Mechanical Turk for workers to "Help researchers understand COVID-19 in America! (Harvard/Northeastern survey, 15-30 minutes)." Workers were paid $6 for completing the survey and the interquartile range for completion time was 19 to 34 with a median of 24 minutes.

Facebook Ads

We recruited volunteers using Facebook ads. We created 32 audiences intended to ensure a diverse panel. These audiences targeted Facebook users in the U.S. by age, income, and gender. (Facebook does not make demographic targeting by race or ethnicity possible.) For each audience, we created two ads recruiting users to participate. We launched ads on September 11th and completed recruitment on September 23rd. We actively managed the budget and ads to minimize the cost per subject and maximize subject diversity, meaning we stopped ads that were expensive and spent more money on audiences that were expensive to compensate for the high cost per subject.

Unlike workers on Mechanical Turk, Facebook volunteers were first asked to sign up for the volunteer panel to receive communications about future waves of data collection. These volunteers were also not compensated for their participation and the interquartile completion time was 22 to 33 minutes with a median of 27.

While the Facebook ads reached a diverse audience by age, gender, and income, the resulting sample was 91% white. In Appendix I, we discuss the potential reasons for this and the results of largely unsuccessful mitigation efforts.

Volunteer Panel

In prior waves, we recruited volunteers using Facebook Ads and organic traffic from the crowdsourcing science website Volunteer Science. These volunteers signed up for the panel before participating in surveys and we email them to return for subsequent waves. The participants are unpaid, and the median completion time was 18 to 31 minutes with a median of 23.

Data Analysis

First, we examine the quality of the data collected from each method. This involved exploring how many participants provided complete data, whether any failed simple attention check questions, and how many provided incredulous data. We remove incomplete data and subjects who fail the attention check. We also remove subjects who provide more than one form of incredulous data.

Our second line of analysis explored the representativeness of the resulting sample. Our primary question was whether the sample was diverse enough to generate nationally representative estimates of public opinion, experiences, and behaviors. We looked at sample diversity on our key weighting criteria - gender, education, age, race, ethnicity, geographic region, and urbanicity. We also examined diversity along relevant but unweighted characteristics like religion, political ideology, party affiliation, income, and employment.

Finally, we weighted the resulting data, compared it to estimates from our national sample, and combined all the data together to create a third estimate. This allows us to directly compare estimates across sampling methods to our current approach using 22,000 subjects from a sample reseller as well as examine the impact of combining the data together. Where possible, we compare all three estimates to similar items from probability-based surveys.

Results

Data Quality

Table 1: Data quality across methods of recruitment

Name	Initial Sample Size	Unusable	Non-Compliant	Suspect data	Final Sample Size
MTurk	1796	4%	3%	4%	1596 (89%)
PureSpectrum	31049	20%	16%	1%	20699 (67%)
Facebook	944	40%	2%	2%	546 (59%)
Recontact	369	27%	2%	1%	262 (71%)

Table 1 summarizes the results of our data quality checks. In our first stage of filtering, we eliminated data that was either incomplete or from the same respondent. 94% of workers provided usable data, by far the highest rate. Recruits from PureSpectrum and our volunteer panel provided roughly the same rate of usable data: 73-80%. Finally, 60% of first-time volunteers from Facebook provided usable data.

We also eliminated respondents who did not pledge to respond authentically and those who failed two attention check questions. Here we find a relatively uniform level of compliance across sampling methods except for PureSpectrum where almost 16% of respondents proved non-compliant as compared to 2-3% for the other methods.

We then evaluated respondent quality based on a few indicators of bogus data: 1) speeding through the survey, 2) selecting the same options, called “satisficing”, and 3) the presence of implausible data. First, for speeding, we compared the distribution of complete times. We used a strict standard of completing the survey in 5 minutes (the average time is 20 minutes) and only one person, a Turker, managed to finish that quickly.

For satisficing, we look at “straightlining:” whether people answer the same response to question after question. We specifically looked at straightlining on all matrix-style questions. Using the ‘careless’ package in R, we looked for the longest sequence of the same answers for each of our matrix questions and added them together. We flagged any data that was more than two standard deviations from the mean — 40 cumulative straightline responses across all questions (the maximum possible was 55).

Finally we look at implausible data: people who provide unrealistic or inconsistent information. Specifically, we flagged people who report having more than 2 religions, living in households larger than 10 people, participating in more than 5 activities outside the household in the past 24 hours, inconsistency in liberal/conservative ideology and democratic/republican party affiliation, and those with inconsistent state and zip codes.

We did not exclude anyone who had just one of these. Instead, for the suspect data filter, we removed anyone who had at least two of these implausible data points or one of these and more than 40 straightline responses.

Sample Characteristics

Each sample is generally diverse on all of the demographics we measured (Table 2). There are substantial numbers of respondents for each demographic group except perhaps religious minorities, participants who identify is non-binary, and participants with less than a high school degree. This extensive diversity means we can generate weighted estimates based on a reasonable amount of cases for each weighted group.

There were important differences between each method, particularly along age, race, education, and employment, and political ideology. PureSpectrum and Facebook offered the most even age distributions (and closest to the national average) while MTurk skewed young and the recontact sample skewed substantially older. The Facebook and the recontact samples were both more educated than the other methods as well as much whiter, and less religious. Mechanical Turk, Facebook, and the recontact sample all leaned more liberal with fewer Republicans than in the PureSpectrum sample.

The national averages in Table 2 allow us to compare each method to a perfectly representative sample. PureSpectrum offered the most nationally representative sample on age, income, education, employment, and race. But it was skewed 70% female and liberal with very few participants without a high school degree. The Mechanical Turk sample provided good coverage by income, education, gender, and race; but oversampled liberals and democrats and those on the lower-middle of the age and income distribution. The Facebook and volunteer panel provided a surprising range as well, but undersampled people with less education, men, the employed, conservatives, republicans, and Christians. Unfortunately, the Facebook and volunteer panel also skewed heavily white - 92-93%. Only 45 respondents reported being non-white, making it very

unlikely that weights could overcome the imbalance and making linear models on race more noisy and prone to spurious results when used in isolation.

Given that the volunteer panel is largely built on Facebook ads, comparing the volunteer panel to the Facebook recruitment 's demographics suggests several selection biases among volunteers. Compared to the Facebook ads, the volunteer panel was older, wealthier, more female, more liberal, and more likely to lean democratic.

Table 2: *Demographics of Participants by Method of Recruitment*

	PureSpectrum	MTurk	Facebook	Volunteer Panel	U.S. Population
Age					
18-24	13	6	8	3	12[+]
25-34	23	41	15	6	18[+]
35-44	22	28	14	13	16[+]
45-55	14	14	19	19	16[+]
55-65	13	8	18	24	17[+]
65+	15	3	21	34	21[+]
Income					
< $25,000	21	16	20	8	17[+]
$25 - $49,999	25	29	24	18	20[+]
$50 - $74,999	20	27	18	19	16[+]
$75 - $99,999	13	17	14	16	12[+]
> $100k	20	11	23	35	34[+]
Education					
Less than HS	3	1	1	0	10[+]
HS Grad	21	10	8	2	28[+]
Some College	34	23	40	26	26[+]
Bachelor Degree	25	51	28	37	22[+]
Graduate Degree	16	16	24	35	3.5[+]
Gender					
Male	30	56	40	32	49[+]
Female	70	43	58	67	51[+]
Genderqueer /Oth.	NA	2	2	1	NR[+]
Employment					
Employed (full)	45	68	34	32	57[++]
Employed (part)	11	9	9	11	--
Retired	16	2	26	34	--

Unemployed	12	5	12	6	8[++]
Gig Worker	1	2	2	2	--
Race/Ethnicity					
White	67	75	92	93	75[+]
Black	13	16	2	1	14[+]
Hispanic	9	6	3	2	19[+]
Asian	6	6	2	1	7[+]
Indigenous	NA	3	3	5	2.1[+]
Religion					
Christian	73	59	46	44	71*
Muslim	1	1	.2	.4	.9*
Buddhist	1	1	3	2	.7*
Nonreligious	21	36	38	40	22*
Hindu	.4	1	1	0	.7*
Jewish	1	1	3	6	2*
Party					
Democratic	34	41	30	49	31[G1]
Republican	30	34	17	9	31[G1]
Independent	30	22	33	33	36[G1]
Other	6	3	19	9	NA
Ideology					
Extreme Lib.	8	14	18	17	24[G2]
Lib.	14	24	22	31	
Slightly Lib.	9	13	10	14	
Moderate	37	18	18	22	35[G2]
Slightly Con.	9	8	9	7	37[G2]
Con.	14	16	15	9	
Extreme Con.	7	7	6	2	
N	20699	1596	546	262	

Standard errors for the PureSpectrum and Mechanical Turk sample were 1% or less. Standard errors for the Facebook sample was 2% or less. Standard errors for the volunteer sample were 3% or less.

[+]Based on 2019 Census estimates

[++]Based on September 2020 Bureau of Labor Statistics estimate

*Based on Pew Religious Landscape study https://www.pewforum.org/religious-landscape-study/

[G1]Based on Gallup 2020 Party Affiliation https://news.gallup.com/poll/15370/party-affiliation.aspx

[G2] Based on Gallup 2019 Report https://news.gallup.com/poll/275792/remained-center-right-ideologically-2019.aspx

Reproducing National Estimates

To answer the question of how these non-probability sample estimates compare to probability-based estimates, we produce a nationally-weighted sample using the PureSpectrum and Mechanical Turk data as well as a combination of all four sources called PS+. We weight based on race, gender, age, geographic region, and urban type (rural/suburban/urban). We compare estimates from all three to those from probability-based surveys across three broad areas (Table 3).

First, we looked at political attitudes around President Trump's handling of the coronavirus, candidate vote choice, and vote choice on a generic congressional ballot. Compared to 538, we consistently under-estimated approval for President Trump's handling of the coronavirus, his support versus Joe Biden, and support for Republican candidates. As we know from the election, the polls themselves also under-estimated support, so our poll was even further off from the election results. Both surveys and their combination also under-estimated disapproval for President Trump compared to the 538 aggregate by six points. The Mechanical Turk panel was within 1 point of 538 in both support for Biden and the generic democratic candidate while the PureSpectrum and combined sample under-predicted democratic support in the generic congressional ballot by 8 and 6 points respectively. This result makes sense given our undersampling of Republicans, simple weighting procedure, and the way we asked these questions.

Second, we looked at less fungible political behaviors: whether they voted in 2016, who they voted for in 2016, and whether they were registered to vote in 2020. The NORC estimated 85% of people were registered to vote. We estimated 89% of the population was registered using the MTurk data and 82% were registered using the PureSpectrum data. Taken together 83% were registered. The vote share for Clinton and Trump in 2016 was 48 to 46. Using MTurk data, we estimated that to be 43 Clinton, 45 Trump and 46 Clinton, 43 Trump using PureSpectrum. Combining them, we estimate a 47:44 Clinton victory. Finally, we look at who voted in 2016 and find substantial variance between the samples. According to MTurk, 77% of eligible voters voted in 2016 and 69% using the PureSpectrum data for a combined average of 74%. In actuality, 58% of eligible voters cast ballots. This over-reporting of actual voting behavior is a common finding in surveys[1].

These results are within 3-5 points from either historical or gold standard data, except for the estimate of voters in 2016 which was off by a substantial margin. This suggests a strong and uncompensated bias towards active voters in both the PureSpectrum and Mechanical Turk samples.

Finally, we looked at Covid and health behaviors and attitudes. Questions here are less 'apples to apples', but there are a few close enough to compare. First, several polls have found that the rate of having been diagnosed with Covid-19 by a medical professional to be 4-6%. We replicate

---

[1] DeBell, Matthew, Jon A. Krosnick, Katie Gera, David S. Yeager, and Michael P. McDonald. 2020. "The Turnout Gap in Surveys: Explanations and Solutions." *Sociological Methods & Research* 49(4):1133–62. doi: 10.1177/0049124118769085.

this using PureSpectrum, showing a rate of 3%. However, the MTurk panel showed a 10% diagnosis rate. This discrepancy however disappears when we compare those who had a positive coronavirus test. In September, the AP-NORC found 4% of the population had had a positive test in September and we find similar rates in the PureSpectrum (3%) and Mechanical Turk (5%) data. In other words, the MTurk panel has an unusually high number of people who were diagnosed without a test.

On the topic of mask compliance, an AXIOS/IPSOS poll found that 88% of people reported they always or sometimes wore a mask when going out. We find a similar rate using a similar question: how often do you follow your state guidance regarding wearing masks in public. Between 89 and 91 percent of people reported following the guidance somewhat or very closely.

Finally, looking at the propensity to get a vaccine, we found that 54% of people were likely and 28% were unlikely to get a vaccine using Pure Spectrum and 67% were likely and 20% were unlikely to get a vaccine using Mechanical Turk. Question wording has varied on this topic. But there were two related questions during this time. The first comes from an NPR/PBS/Marist poll that found 49% of people would get a vaccine and 44% would not. The second from AXIOS/IPSOS found that 64% of people would be likely to get a "safe and effective" vaccine. Both the variations here and the wording differences are substantial, suggesting a clearer question is needed. But these surveys indicate both MTurk and PureSpectrum are likely in the ballpark for how many people would get a vaccine.

Discussion

The PureSpectrum, Mechanical Turk, and Facebook recruitment efforts both capture a diverse pool of largely good-faith participants. However, the Facebook sample was nearly all-white, making it infeasible to produce national estimates. Instead it is more appropriately used as a supplement for the other samples.

Comparing these samples to probability-based samples and ground truth data, we find general agreement among our samples and between our samples and these other sources. There were a couple of notable exceptions. We under-predicted Trump and Republicans' performance in the 2020 election. This was true of polls in general and we did not use a model of likely voters in weighting the estimate.

What is perhaps most instructive is the comparison for whether people would get a vaccine. Here we had two separate probability-based surveys with results that were 15 points away from one another. The supposition is that this difference is due to differences in the way the question is framed. However, between our samples, we found a similar 16-point difference along the same numerical range. Given the consistency in question framing and timing of our surveys, our difference is likely due to sample selection and composition effects. This suggests that attitudes on a potential vaccine vary substantially, which was likely the case in September given there was no vaccine for people to form an opinion of and the process of evaluating candidate vaccines had become politicized.

These results suggest two things. First, the estimates from our non-probability samples conform to those from probability samples. Second, the issues of properly measuring a population with

surveys have just as much to do with the construction, fielding of the survey, and analysis of the data than with whether one uses probability-based or non-probability-based sampling methods.

There is no perfect survey. Even the United States Census struggles with issues about wording and option choices and must make estimates about those who do not respond. What is more useful for making accurate estimates are multiple surveys on the same topic, transparency about data coverage and weighting procedures, and regular evaluations of the effectiveness of the field as a whole. This field-level approach to public opinion requires many more surveys by more diverse research institutions and the only way to do this is through faster, lower-cost, and more scalable non-probability samples.

The bias towards high-cost probability sampling may be increasingly indefensible given the marginal differences in results from lower-cost non-probability designs and the added benefit of fielding more, diverse, reproducible, and transparent surveys rather than relying on single, large, expensive surveys.

Table 3: Comparing Covid States Results to Probability and Administrative Data

Category	Question	Mturk	PS	Mturk + PS	Prob. Samples or Admin. Data
Voter Preferences	*Trump handling of coronavirus*	51% disapprove 33% approve	50% disapprove 34% approve	50% disapprove 34% approve	538.com- 56% disapprove, 40% approve
	Pres. Vote Choice	49% Biden, 38% Trump	47% Biden 39% Trump	48% Biden 38% Trump	538.com- 50% Biden, 43% Trump
	Generic Congressional Ballot	50% Dem 35% Rep	41% Dem 35% Rep	43% Dem 34% Rep	538.com - 49% Dem, 42% Rep
Voter Behaviors	*Voted in 2016*	70% voted	59% voted	64% voted	Official total: 58% voted
	Vote Share in 2016	45% Trump, 43% Clinton	44% Trump, 46% Clinton	44% Trump, 47% Clinton	Official total: 46% Trump, 48% Clinton
	Register'd in 2016	89%	82%	83%	AP/NORC: 85%
Covid-19	*Diagnosed with C19 by a medical professional*	10%	3%	4%	USC: 4%; AP/NORC: 6%
	Had a test come back positive	5%	3%	4%	AP-NORC - 4%
	Num Days waiting for test results	3.8	3.5	3.6	None
	Would get a vaccine	67% likely, 20% unlikely	54% likely, 28% unlikely	57% likely, 26% unlikely	64% "likely to get a safe and effective vaccine" (ABC/IPSOS), 49% "will get a vaccine," 44% will not (NPR/PBS/Marist)
	Mask compliance	89% very or somewhat closely	91% very or somewhat closely	91% very or somewhat closely	AXIOS/IPSOS 88% always or sometimes wear mask when out

Note: Comparisons are made from surveys that were fielded with the same month as the Covid States. Surveys and items were found through the Societal Experts Action Network (SEAN) which archives surveys related to Covid-19 at https://covid-19.parc.us.com

Appendix I

Recruiting Racial and Ethnic Minority Participants from Facebook

In August 2020, Facebook removed advertisers' ability to target ads based on race and ethnicity. These restrictions are meant to prevent advertising that perpetuates discriminatory commercial practices and extends to prohibitions on targeting by religion and sexual orientation (though not gender).

However, this does not mean that Facebook's algorithms are blind to race or ethnicity or other historically marginalized social statuses. In our attempt to recruit participants for our survey in September, 92% of our respondents reported being white while the U.S. population is 75% white. Our audiences were constructed only using age, gender, and income demographic filters. There were no interests or behavior targets. And, while research has shown that Facebook's ad algorithm has used image content to predict who will respond to an ad[2], we only used two images, one which featured two White people and one an illustration of the coronavirus.

In November, we followed up with small-scale ad tests, using custom audiences, geo-targeting, images of racial and ethnic minorities, and ethnic interest targets to try and reach these participants. In our first test, we targeted ads to four audiences, a lookalike audience of non-white participants from Volunteer Science and three custom audiences composed of the zip codes with the highest percentage of Latinos, indigenous people, and African Americans. For each audience, we used advertising images that included members of that racial or ethnic minority group. These ads failed, resulting in a sample that was still 92% White.

We tried a second wave of experiments using interest-based targeting. We used interest-based targeting to create audiences for each racial and ethnic group and, again, matched images to audiences by race and ethnicity. The results of these targeting efforts were samples that were 85% white.

---

[2] Ali, Muhammad, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. "Discrimination Through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes." *Proc. ACM Hum.-Comput. Interact.* 3(CSCW):199:1-199:30. doi: 10.1145/3359301.