The European Political Science Association, 2016 *doi:10.1017/psrm.2016.XX*

Survey Sampling in the Global South Using Facebook Advertisements *†

LEAH R. ROSENZWEIG, PARRISH BERGQUIST, KATHERINE HOFFMANN PHAM, FRANCESCO RAMPAZZO, AND MATTO MILDENBERGER

Survey research in the Global South has traditionally required large budgets and lengthy fieldwork. The expansion of digital connectivity presents an opportunity for researchers to engage global subject pools and study settings where in-person contact is challenging. This paper evaluates Facebook advertisements as a tool to recruit diverse survey samples in the Global South. Using Facebook's advertising platform we quota-sample respondents in Mexico, Kenya, and Indonesia and assess how well these samples perform on a range of survey indicators, identify sources of bias, replicate a canonical experiment, and highlight trade-offs for researchers to consider. This method can quickly and cheaply recruit respondents, but these samples tend to be more educated than corresponding national populations. Weighting ameliorates sample imbalances. This method generates comparable data to a commercial online sample for a fraction of the cost. Our analysis demonstrates the potential of Facebook advertisements to cost-effectively conduct research in diverse settings.

*Leah R. Rosenzweig is a Senior Fellow at the Center for Global Development (lrosenzweig@cgdev.org). Parrish Bergquist is Assistant Professor of Political Science, University of Pennsylvania (pberg@upenn.edu). Katherine Hoffmann Pham is AI Advisor at UNICEF's Office of Innovation (note: this work was conducted independently of UNICEF and does not represent the views or interests of UNICEF or any other UN entity). Francesco Rampazzo is Lecturer in Demography, Department of Sociology, Leverhulme Centre for Demographic Science, and Nuffield College, University of Oxford. Matto Mildenberger is Associate Professor of Political Science, University of California Santa Barbara. The authors Chris Bail, Matt Salganik, Anne Helby Petersen, Julien Migozzi, and Tina Law for their assistance in shaping this project. We thank Michaël Aklin, André Grow, Peter Howe, Anthony Leiserowitz, Jennifer Marlon, Umberto Mignozzetti, Blair Read, Baobao Zhang, Ingmar Weber, Alessandro Sorichetta, Dennis Feehan, Chris Tausanovitch, researchers at the Busara Center for Behavioral Economics, Afrobarometer, and seminar participants at NYU Abu Dhabi for helpful comments. Thanks to Emma Franzblau, Gabriel de Roche, and Ingmar Sturm for research assistance. We thank Warsama Abdifitah, Kibuchi Eliud, Ahmed Hared, Lilian Ligeyo, Laban Okune, Gustavo Ovando-Montejo, Nelson Ngige, and Eunice Williams for translation assistance. This work was supported by the Summer Institute for Computational Social Science (SICSS), the Russell Sage Foundation, the Alfred P. Sloan Foundation, the Yale Program on Climate Change Communication, and Meta. To view supplementary material for this article, please visit http://dx.doi.org/10.1017/psrm.2014.11. Competing interests: The authors have none to declare.

†First two authors listed are joint first authors whose names have been randomly ordered. Bergquist is the corresponding author. Contributing co-authors' names are listed in random order.

Word count: 9,037

Survey research in the Global South traditionally requires local enumerators to conduct face-to-face surveys with respondents, necessitating large budgets and lengthy fieldwork. Historically, there have been few alternatives to in-person recruitment due to poor electricity coverage and limited phone or internet connectivity. However, today much of the world's population is digitally accessible. By the end of 2022, an estimated 68% of the world's population had a mobile phone subscription, and approximately 55% had access to mobile internet (GSMA 2023; Rotondi et al. 2020). This connectivity presents an opportunity for researchers with limited budgets and those seeking to collect data in settings where the resource-intensive models of research are challenging, such as due to natural disasters, violent conflicts, or pandemics.

This paper evaluates one emerging method to recruit online samples: Facebook advertisements. The analysis builds on prior work finding that low-cost, online platforms allow scholars to generate valid survey-experimental results (Berinsky, Huber, and Lenz 2012; Coppock and McClellan 2019; Mullinix et al. 2015). We take this research program a step further, asking if Facebook can be used to generate high-quality samples for broader survey research purposes. According to Meta, Facebook's parent company, Facebook had over 3 billion monthly active users as of June 2023, (Meta 2023), more than a third of the global population. As such, it is the most widely used social media platform around the world (Ortiz-Espina 2019). Given this massive user base, the platform offers researchers potential access to nationally, culturally, and demographically diverse global populations. Scholars are already using the platform to quickly and cheaply recruit diverse populations relative to other modes of survey recruitment (Ramo et al. 2014; Kapp, Peters, and Oliver 2013; Grewal 2023; Jäger 2017; Pötzschke and Braun 2017; Hirano et al. 2015; Rosenzweig 2021; Samuels and Zucco Jr 2014; Noh, Grewal, and Kilavuz 2023; Kilavuz, Grewal, and Kubinec 2023; Finkel, Neundorf, and Rascon Ramirez 2023; Broockman

and Green 2014; Ryan 2012; Bond and Messing 2015; Grow et al. 2020; Bicalho, Platas, and Rosenzweig 2020; Offer-Westort, Rosenzweig, and Athey 2024; Jacobson 2024). Yet, survey researchers caution against an uncritical adoption of new recruitment methods without considering possible limitations on survey sample quality (eg Ansolabehere and Schaffner 2010; Berinsky, Huber, and Lenz 2012; Ternovski and Orr 2022). This paper provides a systematic assessment of the opportunities and drawbacks of Facebook advertisements as a recruitment tool in the Global South.

While most research has used Facebook to recruit convenience samples or specific target populations, an emerging literature has assessed if the platform can also recruit nationally representative samples. As Boas, Christenson, and Glick (2020) demonstrate in India and the United States, Facebook's high penetration and diverse user base make the platform an attractive opportunity for recruiting nationally representative samples. Zhang et al. (2020) demonstrate this potential by using targeted Facebook advertising to cost-effectively recover an approximately nationally representative sample in the United States. Neundorf and Öztürk (2023) build on this work by showing how different targeting strategies influence the cost and representativeness of samples recruited in the UK, Turkey, Spain, and Czechia. This research underscores Facebook's potential for survey sampling, but researchers are still left with a geographically and conceptually incomplete understanding of the platform's benefits and drawbacks.

Here, we expand the geographic scope and conceptual underpinnings of Facebook's potential as a recruitment tool for survey respondents. We follow prior work (Berinsky, Huber, and Lenz 2012) by using a canonical survey experiment to show that Facebook passes the relatively low bar of providing a platform for conducting valid survey experimental research. We then explore the tool's use for more general public opinion research. We compare the demographic composition and estimates of political attitudes from Facebook-

recruited samples in Kenya (n=1,528), Mexico (n=5,168), and Indonesia (n=3,277) with sample composition and public opinion estimates derived from nationally representative benchmark data sets in each country.[1] We also compare our Facebook sample in Indonesia with an online sample recruited by a professional survey firm, which is the likely alternative for most researchers. To provide the conceptual scaffolding for our analysis, we apply the Total Survey Error framework (Deming 1944; Groves and Lyberg 2010; Ansolabehere and Schaffner 2010). Our analysis of survey error highlights the sources of potential bias in Facebook survey samples, exposes which sources of bias researchers can control, and illuminates how researchers can reduce bias. Whereas most existing studies have focussed on the U.S. and Europe, we evaluate the representativeness of Facebook-recruited samples in low and middle-income countries that have lower levels of internet access, literacy, and Facebook marketing investment.

We show that quota sampling, which is also useful in other survey modalities, can help to overcome bias introduced by using the Facebook user base as a sampling frame for national populations. Given the noisiness of the ad platform's underlying demographic data, we highlight the importance of applying survey weights based on self-reported demographic data to ameliorate representational biases such as, in our case, the over-representation of highly educated individuals. We then examine the political behaviors and attitudes reported by our Facebook-recruited samples. Facebook-recruited individuals report being more politically engaged and more supportive of environmental protection than those recruited by in-person surveys. These differences may be due in part to survey mode and in part to sample composition. Practically, we find that the Facebook platform

[1]We filed two pre-analysis plans for this analysis that are available [link removed for anonymous peer review] and [link removed for anonymous peer review].

allows for quick and cheap recruitment. Our surveys took one to three weeks to field, and cost an average of $1.03 per completed survey (ranging from $0.17 to $1.57). Taken together, our findings point to the potential of Facebook-recruited samples in helping researchers access diverse communities across the Global South. Researchers should tailor their use of this sampling method to both the type of research question asked and the target population of interest, and we show specific steps that researchers can take to reduce survey bias. More broadly, many other survey recruitment and implementation modalities face challenges of bias, noise, and non-response that are similar to those we highlight with Facebook. As such, we hope this paper is broadly useful to researchers considering the appropriateness and tradeoffs involved in using other new social science research platforms.

Conceptual framework: Defining and measuring sources of survey error in public opinion samples

We begin by developing a definition of survey quality building from assessments of cost-quality tradeoffs in public opinion research. Scholars have assessed the quality of samples recruited through Amazon's Mechanical Turk (MTurk) (Berinsky, Huber, and Lenz 2012; Huff and Tingley 2015), Prime Panels (Litman, Robinson, and Abberbock 2017), Lucid Fulcrum Exchange (Coppock and McClellan 2019; Ternovski and Orr 2022), Google Consumer Surveys (Santoso, Stein, and Stevenson 2016), and Facebook Advertising (Neundorf and Öztürk 2023; Kosinski et al. 2015; Jäger 2017; Boas, Christenson, and Glick 2020; Zhang et al. 2020). Much of this work focuses on the quality of U.S. online samples, typically by comparing sample demographic statistics with the U.S. Census (Berinsky, Huber, and Lenz 2012; Huff and Tingley 2015; Coppock and McClellan 2019;

Zhang et al. 2020). These assessments are relatively straightforward in the US, where high-quality census data are frequently updated and many probability and quota samples are available as benchmarks.

Extending this work, we conceptually disaggregate and empirically assess the types of errors that threaten the validity of conclusions drawn from Facebook surveys. We define error as unobserved disturbances that influence a statistical quantity of interest. Such error is problematic if it causes survey estimates to differ systematically from true population parameters. These systematic differences indicate bias in our estimates.

We use the Total Survey Error framework, first developed in the 1940's (Deming 1944) and used by modern survey researchers (eg, Groves and Lyberg 2010; Groves et al. 2011; Lyberg and Weisberg 2016; Ansolabehere and Schaffner 2010), to define the distinct sources of error that threaten the external validity of conclusions derived from Facebook-recruited surveys. Since the framework's introduction, survey methodologists have defined two groups of quality indicators for survey statistics: measurement error[2] and representation error (Groves et al. 2011; Groves and Lyberg 2010). Here we focus on the latter. Errors of representation are systematic or random imperfections in the relationship between a target population, sampling frame, and sampling units, and they threaten external validity.

Figure 1, adapted from Groves et al. (2011) and Groves and Lyberg (2010), shows the inferential steps required to draw population conclusions from a Facebook-recruited

[2]Measurement error addresses gaps between the constructs that a survey is designed to assess, the measures used in a survey, and responses to those measures. Measurement error is primarily a matter of questionnaire design and is not platform-specific. Still, we do examine one type–respondent attention–in the SI to assess if this component of measurement error might affect Facebook sample quality differently than on other platforms.

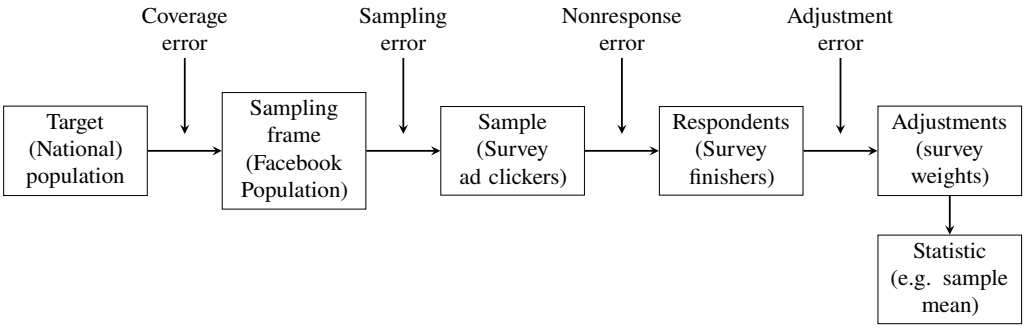


Figure 1. Components of Representation Error in Facebook samples

survey sample. First, coverage error is defined as the gap between a target population and a sampling frame. In our case the target population is a country's national adult population, and the sampling frame is the national adult population with a Facebook account in that country. The existence of adult residents of each country without Facebook will generate "undercoverage." Coverage bias is quantified as the difference between the mean value of a descriptive statistic in the national population and the Facebook population. Since we cannot directly observe population-level average political attitudes, we instead infer this quantity from nationally representative, in-person benchmark surveys.

Second, sampling error arises because not all individuals in the sampling frame are surveyed. Random sampling variance occurs since many different sets of individuals could be drawn from the sampling frame, simply by chance. Sampling bias arises when individuals in the sampling frame have different chances of being included in the sample. Third, unit non-response error arises when some sampled individuals fail to record high-quality, complete survey responses. Unit non-response is the gap between people who click on the advertisement for our survey and those who finish the survey. Non-response bias arises if certain types of people are more likely to finish the survey than others, and if there is a relationship between the likelihood of finishing a survey and

survey responses. Finally, adjustment error arises when researchers weight data to give greater representation to cases that are under-represented in the sample. These weights are used to reduce coverage, sampling, and non-response bias, but they can also increase these biases (Bailey 2024).

We examine total survey error and disaggregate its components with two sets of empirical analyses. Our analyses build on prior work that compares survey-derived statistics to external benchmarks, such as a "gold-standard" survey, census, or re-interview data (Groves and Lyberg 2010; Berinsky, Huber, and Lenz 2012; Coppock and McClellan 2019; Zhang et al. 2020; Ansolabehere and Schaffner 2010; Pew Research Center 2018; Holliday et al. 2021). We compare, first, the demographic compositions of our samples and, second, the public opinions measured in our surveys with those derived from other high-quality and commonly used samples. Our disaggregated assessment of survey errors and their associated biases allows researchers to gauge Facebook's suitability for specific research applications and improves our understanding of error types under researchers' control.

Data Collection

Case selection

We fielded surveys in Mexico, Kenya, and Indonesia. We selected case countries from three different continents that are neither best nor worst cases in terms of Facebook usage, where recent and accurate census data is available, and where high literacy rates and mobile phone access ensure broad accessibility of an online survey. In Mexico, Facebook penetration, as a percent of the adult population, is 87% and comparable to other countries

in the region, such as Brazil (72%) and Argentina (83%).[3] In Kenya, Facebook penetration is 25%, compared to 45% in South Africa, 22% in Nigeria, and 13% in neighboring Tanzania. In Indonesia, Facebook penetration is 76%, compared with 94% in Malaysia, 92% in Singapore, 60% in India, and 58% in Laos.[4] Our three case countries also have high rates of mobile phone use, with 100, 122, and 115 cellular subscriptions per 100 people in 2022 in Mexico, Kenya, and Indonesia, respectively.[5] This is crucial since most people use their mobile phones to access Facebook. Finally, adult literacy is high in all three countries (95% in Mexico, 82% in Kenya, and 96% in Indonesia), allowing the majority of citizens to read and self-administer online surveys (World Bank 2020).

Benchmark data sets

We compare Facebook-recruited samples against high-quality data benchmarks in each case country. First, we use national censuses. Second, we use well-respected, in-person, nationally representative surveys: the Latin American Public Opinion Project (LAPOP) Americas Barometer, the Afrobarometer, and the Asian Barometer. In Indonesia, we also compare our Facebook sample with an original survey we fielded with Dynata, a survey firm that recruits respondents through its online panel. We make the comparison with Dynata to provide insights about the comparative quality of Facebook samples with the most viable alternative for many researchers considering online surveys. We present the dates of data collection for each dataset in Table 1.

[3]Calculated based on Facebook-reported users per country (Araujo et al. 2017) and UN-derived population figures for individuals age 15 years and older in each country (United Nations 2019; Ševčíková 2020).

[4]https://www.internetworldstats.com/stats3.htm#asia.

[5]https://data.worldbank.org/indicator/IT.CEL.SETS.P2?locations=MX.

Quota sampling

To minimize sampling bias, we use a stratified sampling approach designed to mimic the demographic-geographic stratified sampling approaches used by our in-person comparison benchmark surveys (LAPOP, Afrobarometer, and Asian Barometer). For the LAPOP and Afrobarometer benchmarks, we designed Facebook geographic strata to approximate the benchmark survey sampling approaches as closely as possible. For Indonesia, we used a stratified sample by province. Within each geographic stratum, we designed target cells based on the demographic characteristics used in our benchmark in-person surveys: gender in Kenya and both gender and age in Mexico and Indonesia. We then attempted to correct observed or expected imbalances by targeting additional respondents within underrepresented categories, including education (in Mexico and Kenya) and age (in Kenya).

Facebook allows for two different types of geographic targeting. Researchers can directly target audiences by providing a point of interest (either an address or a set of latitude and longitude coordinates), as well as a radius defining the catchment area. In Kenya, we used this approach.[6] Alternatively, researchers can use Facebook's predefined geographic entities, which typically consist of neighborhoods, cities, or sub-national administrative units. In Mexico and Indonesia, we targeted respondents using Facebook's predefined geographic units: municipalities in Mexico and provinces in Indonesia. Table 1 summarizes the sampling strategy for each country. The SI further details each country's sampling approach and discusses the constraints of Facebook's advertising platform.

[6]We used the Afrobarometer sampling clusters geolocation.

TABLE 1 *Comparison of data collected in Mexico, Kenya, and Indonesia*

	Mexico	**Kenya**	**Indonesia**
Demographic targeting	Gender, age, (education)*	Gender, (age, education)*	Gender, age
Geographic targeting	Administrative unit	Grouped Afrobarometer clusters	Administrative unit
Field dates	Aug 17-Sept 10, 2019	Sept 21-29, 2019	July 5-18, 2023 Oct 31-Nov 16, 2023
Incentives	No	Yes (~ $0.50)	Yes (~ $0.65)
Comparison data sets	2015 Census 2019 LAPOP (Round 8)	2019 Census 2019 Afrobarometer (Round 8)	2020 Census 2019 Asian Barometer (Wave 5) 2021 Dynata
Question types	Demographics, political party affiliation, climate change beliefs	Demographics, mobility, political attitudes, social media use, fertility, household assets, behavioral experiment	Demographics, political attitudes, climate change beliefs, social media use, household assets, behavioral experiment
N. questions	23	60	53
N. quota sampling cells	128	66	272
N. respondents	5,168	1,528	3,277

Note: * Parentheses denote strata added partway through data collection to correct for observed sample imbalances.

Survey instruments

To direct respondents to the surveys, we created Facebook pages representing our survey campaigns, and placed ads from these pages targeting each sampling strata. After clicking on a Facebook ad, respondents were sent to a Qualtrics-hosted survey. Respondents provided informed consent before agreeing to participate in the survey. In Mexico the survey was administered in Spanish. In Kenya, the first survey question asked respondents to choose from one of five possible languages (English, Kiswahili, Kikuyu, Luo, and Somali). The Indonesian survey was offered in Bahasa. After completing the survey, respondents were directed to a thank-you page with an embedded Facebook "pixel" which

allowed Facebook to identify users who completed the survey.[7]

In all three surveys we collected information on demographics and attitudes, which we used to compare Facebook samples against national census and benchmark survey data. The Kenya survey replicated questions from the Kenyan Census and the Afrobarometer survey. It was fielded immediately following the 2019 Census and concurrent with the 2019 Afrobarometer. The Mexico survey replicated certain questions from the Mexican Census and the LAPOP survey, fielded in early 2019 (LAPOP 2018–2019). The Indonesian survey replicated survey questions from the Asian Barometer and an Indonesia survey sample recruited by Dynata in October 2021. Full copies of our surveys are included in our Harvard Dataverse replication archive.

Survey weights

For all samples, we used iterative proportional fitting, or raking, to create weights for respondents who completed the survey. Our weights are designed to reflect the distribution of the national populations (as measured in the national census) across gender, education, age cohort, and geography. We used the marginal, rather than joint, distributions, to avoid excessively over-weighting respondents with very rare combinations of values on the raking variables. We also created an upper bound for the weights at the 95th percentile of the original distribution, to avoid excessively over-weighting very rare respondent types. Full details are included in the SI.

[7]The Facebook pixel did not successfully register completes in the Indonesia survey. This could lead to more efficient ad spending/higher conversions, but might also imply a slightly different sample composition. We found similar results in terms of over-represented subgroups in all samples, including in the Kenya and Mexico samples that did have a succesfull pixel integration.

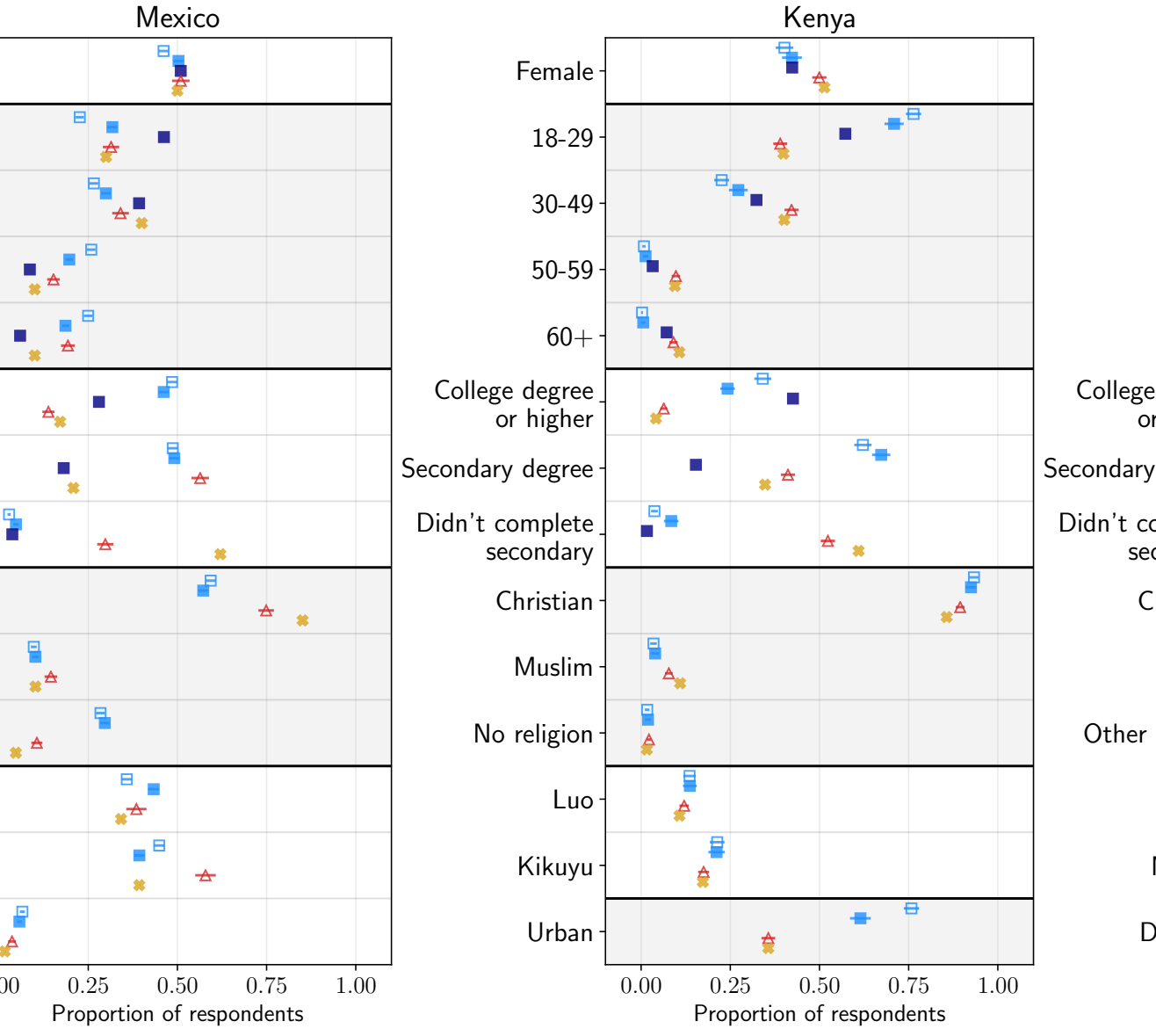
Comparing Demographics

To evaluate the quality of the statistical summaries derived from Facebook samples, we first examine total survey error by comparing the demographic characteristics of our samples with benchmark surveys and national census data. Our quantity of interest is the weighted sample mean survey response.

Figure 2 plots the distribution of demographic characteristics, compared with benchmark surveys and national census data. In all three countries, the weighted Facebook samples (filled, dark blue squares) differ most from the census (gold crosses) with respect to age and education. In Kenya and Indonesia, Facebook survey samples are younger than the national population and the barometer samples (red triangles). By contrast, the Mexican Facebook sample contains a greater share of respondents 50 years and older than the national population, similar to the LAPOP bias.

In general, Facebook samples are more educated than national populations and the barometer surveys. However, the Indonesian Facebook sample is less biased on education than the commercial online sample (the most viable alternative for most researchers). In Indonesia 8% of citizens have a college degree, as reported in the census, compared with 22% of respondents in the weighted Facebook sample and 63% of respondents in the Dynata sample (purple circles). The commercial online sample also under-represents lower education respondents even more severely than our Facebook sample.

We find little gender bias across all three countries. The distribution of religions in Kenya and Indonesia is also fairly representative, although some bias is evident in Mexico. Marital status in Mexico and Indonesia, and tribe in Kenya, also show minimal bias. However, there is substantial bias towards urban populations in Kenya. In the weighted Facebook sample, about 60% of respondents report living in a mostly or completely urban


Mexico
Kenya
Female
18-29
30-49
50-59
60+
College degree or higher
Secondary degree
Didn't complete secondary
Christian
Muslim
No religion
Luo
Kikuyu
Urban
0.00
0.25
0.50
0.75
1.00
Proportion of respondents
Facebook Sample
Facebook Sample (weighted)
Facebook Population
Dynata
Dynata (weighted)
LAPOP/Afrobarometer/As

context - compared to 36% reported in the census.

Coverage error

To examine the specific sources of these biases, we begin with coverage bias: the mismatch between a target population and sampling frame. Here, we compare the national population (described by the census and shown as gold crosses in Figure 2) with the Facebook population (filled, dark blue squares in Figure 2) in each country. Around the time of data collection for each Facebook survey, we used `pySocialWatcher`, a Python package, to retrieve age, gender, and education data for Daily Active Users and Monthly Active Users from the Facebook API (Araujo et al. 2017). These data estimate the demographic characteristics of the full Facebook population.

Coverage bias seems to contribute to the over-representation of more highly educated respondents observed in our Facebook samples. In all three countries, the Facebook population severely under-represents individuals who did not complete secondary school, as do all of our Facebook samples. Similarly, those with a college degree are over-represented in the Facebook populations. In Kenya, education-related coverage bias is even larger than overall bias (the difference between the census and our weighted Facebook estimates), and quota sampling appears to alleviate this bias. In Mexico, by contrast, college-educated individuals are more over-represented in the Facebook sample than in the Facebook population. Thus while coverage bias may account for some of the over-representation of more educated respondents in our samples, sampling and non-response error are also contributing factors, as discussed below.

Coverage bias also helps account for the over-representation of young people in the Kenya and Indonesia Facebook samples. The Facebook population in both countries

over-represents 18-29 year-olds and slightly under-represents older individuals. Still, the Kenyan and Indonesian Facebook samples over-represent 18-29 year-olds even compared with the Facebook population. This suggests that coverage bias only partially accounts for the over-representation of young respondents.

By contrast, coverage bias cannot account for the over-representation of older individuals in the Mexico Facebook sample. Similar to the cases of Indonesia and Kenya, the sampling frame in Mexico over-represents young people, but our unweighted Facebook sample approaches the census proportion of young people. We attribute this better balance to our use of more fine-grained quotas in Mexico. Our stratified sampling approach helped us overcome coverage bias in recruiting our sample in Mexico.

With respect to gender, the Facebook population matches the census in Mexico quite closely, but in Kenya and Indonesia the Facebook populations under-represent females compared to the national populations. The Kenya Facebook sample (weighted and unweighted) mirrors this under-representation. In Indonesia, however, the Facebook sample (weighted and unweighted) closely approximates the gender balance in the national population.

To assess how much of an issue coverage bias might generate, researchers can use the Facebook Marketing API to examine the Facebook population data for their target population of interest before beginning data collection. Stratified quota-based sampling and population-based weighting can help mitigate coverage bias even in cases where the sampling frame may not be representative of the population on a particular dimension.

Sampling error

Quota-based sampling can help researchers mitigate coverage bias, but we need to examine whether the design of the Facebook ad platform limits researchers' ability to target ads towards particular subgroups. Sampling error arises because only some of the individuals in the sampling frame (Facebook users in each country) are included in the potential sample (Facebook users who are shown the ad and click on it). Sampling bias could arise at two points in the sampling process. First, ad design may appeal to some individuals more than others. Researchers have control over this step. They can, for example, experiment with the ad design to appeal to different respondent types. Second, the Facebook advertising platform may systematically fail to reach some individuals. Researchers can target certain individuals with recruitment quotas, but whether ads actually reach targeted individuals is determined by Facebook's back-end data and algorithms, which are beyond researchers' control.

To assess sampling bias, we focus on individuals who clicked our survey ad ("ad clickers"). The Facebook ad platform allows us to observe some demographic information for ad clickers, even if they did not complete the survey, because the survey records the sampling stratum of the ad they clicked on. We use this Facebook-inferred demographic information to check whether we recruited at least one individual from each stratum and, conversely, whether any individual types were systematically excluded. In Kenya and Mexico, we received responses from every quota, although in Kenya we did not reach quota targts in 49 of 66 strata.[8] In Indonesia, we were unable to recruit respondents from 67 (25%) of 272 targeted strata. The vast majority of these strata contained men and

[8]Strata targets were set according to the Afrobarometer's population-derived weights associated with each stratum's geolocation.

women over 50 years old, though we also failed to reach women between 30 and 49 years in three provinces. Thus, in Indonesia, sampling bias contributes to the under-representation of older individuals in our sample.

Of course, if Facebook's back-end data are inaccurate, then even perfect success at recruiting people from survey strata will not ensure that all individuals have a chance of being sampled. To examine this possibility, we compare self-reported demographics with those reported by Facebook. Table 2 provides the percentage of people in each category for which self-reported and Facebook-targeted characteristics match.

TABLE 2 *Accuracy of Facebook targeting, as defined by the percent match between Facebook- and self-reported data*

Characteristic	**Mexico**	**Kenya**	**Indonesia**
Primary sampling characteristics:			
Gender	99%	91%	76%
Age	87%	44%*	77%
Location⋄	67%	73%	55%
Additional criteria:			
Education level	70%	12%‡	

Note: * This reflects the proportion of respondents who were at least 32 years old, given that they responded to an ad targeting this age group. The ages of these respondents ranged from 19 to 48 years old, with a mean of 31 years.
‡ This reflects the proportion of respondents who reported some secondary school or less, given that they responded to an ad targeting respondents with an "unspecified" level of education.
⋄ For location matches we aggregate self-reported and targeted location to the largest administrative unit used for targeting. In Kenya this corresponds to the former provinces used as administrative units, matched from the county respondents say they live in and the locations used for ad targeting. For Indonesian respondents we checked whether respondents self-reported living in the same province as the targeted province for the ad through which they were recruited. For Mexican respondents we checked whether respondents self-reported living in the same group of municipalities that the ad through which they were recruited was set to target. The "Recruitment and sampling" section of the SI provides additional clarity on the administrative units used for sampling in each country.

The accuracy of Facebook's ad targeting varies across demographics and between countries. Facebook's targeting was remarkably accurate in correctly identifying respondents' gender in Mexico and Kenya (99% and 91% match, respectively), and slightly less

accurate in Indonesia (76%). In both Indonesia and Kenya, the 10-20% of respondents recruited from an ad targeting the opposite gender might have resulted from respondents sharing ads with friends. It would not be surprising that greater sharing would have occurred in the context of the incentivized surveys in Kenya and Indonesia, compared to the non-incentivized survey in Mexico. Targeting by age was quite accurate in Mexico (87% match) and Indonesia (77% match), but not in Kenya (44%). Geographic targeting was similarly accurate in Mexico and Kenya but less so in Indonesia. Furthermore, in Kenya, education targeting was very imprecise, which is in part due to the fact that the targeting in Kenya was focused on those with an unspecified education – meaning they had not reported their level of education on their profile. By contrast, education targeting in Mexico was reasonably accurate. We targeted respondents with no higher than a high school degree, and 70% of our respondents self-reported an education level consistent with this targeting criterion. We report more details on these comparisons in the SI.

These findings indicate that inferred demographics are not always accurate, highlighting the importance of collecting self-reported demographic data. We also recommend researchers periodically examine the composition of the sample of ad clickers to gauge success at recruiting respondents from each quota. As an example, SI Figure S1 shows the proportion of age and gender groups (as inferred by Facebook) for those who clicked on ads in each country. Researchers can use comparisons like those shown in Table 2 and Figure S1 to iterate their targeting strategy while their survey is in the field. To increase entrants from under-represented groups in the sample of ad clickers, researchers could increase spending for these groups' ad sets, modify ads to increase appeal for these groups, or leave target ads running for a longer period.

Overall, quota sampling helps us recruit a more diverse sample of respondents. Left to its own devices, Facebook's algorithms would optimize ad targeting to recruit the least

expensive sample. This optimization entails recruiting respondents that are most similar to those who have already entered the survey. Running an ad campaign across quota sampling cells weakens this regression towards the most prevalent (or cheapest) respondent types. Still, the errors in Facebook's back-end estimation of demographic attributes, as well as the fact that some people are excluded from the Facebook platform altogether, means that researchers cannot fully overcome the gap between the Facebook and national populations through quota sampling alone.

Non-response error

Unit non-response is common across survey contexts, and causes bias when non-random and correlated with survey items (Bailey 2024). We define attrition and non-response interchangeably, as entering the survey but failing to provide high-quality and complete responses. Overall, the attrition rate was 46%, 31%, and 25% in Mexico, Indonesia, and Kenya, respectively. For context, attrition from the online sample for the American National Election Study was 14% in 2020 (DeBell et al. 2020, pg. 74), and attrition from the Cooperative Election Study was 21% in 2022 (Ansolabehere, Schaffner, and Shih 2023, pg. 11).[9] We examine whether non-response is systematic using the same Facebook-assigned demographic information used in the previous section. We first assess whether the likelihood of entering but dropping out of the survey was correlated with users' demographic characteristics (assigned by Facebook). Figure 3 shows that attrition is systematic, but the predictors of attrition vary across samples.

[9]We would expect attrition from these canonical surveys to be lower than ours because the samples consist, wholly or partially, of empaneled respondents whose compensation for participating ($40-$200 in the case of the ANES) is substantially higher than the airtime we provided (ANES methodology report, page 29).

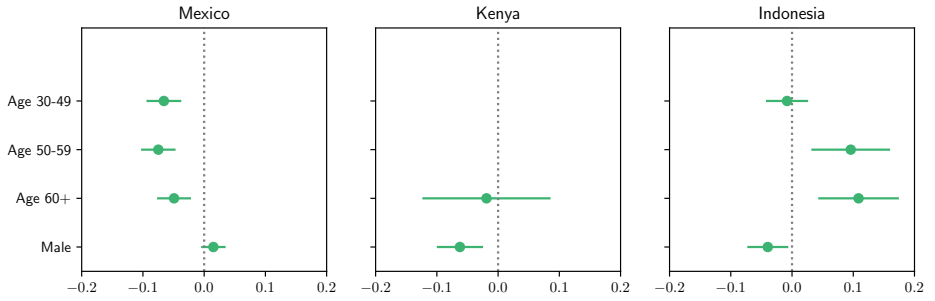


Figure 3. Predictors of non-response
Note: Figure shows the results from a linear regression of attrition (attrition=1, completion=0) on demographic characteristics used by Facebook to target individuals. Omitted categories are Female and 18-29 for Mexican regressions, Female and 21-29 for the Indonesian regressions, and Female and younger than 32 years old for the Kenyan regressions. 95% confidence intervals are reported using heteroskedasticity-robust standard errors.

Incorporating age, gender, and education into adjustment weights helps to alleviate non-response bias associated with these observable characteristics. We reduce non-response bias by upwardly weighting individuals with under-represented characteristics and down-weighting those with over-represented characteristics.[10] Researchers could alternatively adjust their sampling strategy on the fly to improve response rates among high-attrition groups. For instance, researchers might increase ad spending or introduce incentives for high-attrition groups. Of course, unobserved sources of non-response present a more worrisome threat to descriptive inference, since researchers cannot adjust for unobserved and non-ignorable sources of non-response. To account for non-ignorable non-response, scholars can use bounds (Manski 1990), sensitivity analysis (Hartman and Huang 2024), selection models (Gomes et al. 2019; McGovern, Canning, and Bärnighausen 2018),

[10]Note that this correction is imperfect, since our analysis of non-response bias is based on Facebook-assigned demographics, whereas our weights are based on self-reported demographics.

non-response weights (Sun et al. 2018), or other methods (Bailey 2024).

Adjustment Error

While we can minimize unit non-response by weighting on observable characteristics that predict non-response, weights could also exacerbate other biases. This would reflect adjustment bias and occurs when responses within any weighting strata are unrepresentative of public attitudes within that strata. More specifically, adjustment bias occurs if, within groups defined by the weighting variables, non-response is non-ignorable (ie, correlated with unmeasured characteristics that are also correlated with the outcomes of interest) (Bailey 2024). If this is the case, upweighting or downweighting these unrepresentative individuals will introduce bias into the sample average. We examine adjustment error by first examining whether weights correct demographic imbalances in the sample. We find that our (trimmed) weights do ameliorate biases. We then investigate whether these weights introduce bias on other variables we measure and for which we have benchmark comparisons, and find they do not.

Figure 2 summarizes how the application of weights corrects demographic imbalances by comparing unweighted (hollow, light blue squares) and weighted (filled, light blue squares) Facebook samples with census data (gold crosses). Weighting corrects the slight gender imbalances in all three samples. For age and education, weighting reduces the distance between the Facebook samples and the national populations but does not completely eliminate it. In part, the remaining imbalances in age and education are due to trimming weights at the 95th percentile, to avoid excessively over-weighting very rare respondent types. The Facebook sample almost perfectly matches the census populations when weights are not restricted.

While weighting improves the representativeness of the sample on the demographic characteristics incorporated into the weights, it could introduce bias if individuals within our sample reflect a skewed sample of the groups that we weight them to represent. We diagnose this possibility by investigating the impact of weighting on the accuracy of descriptive inferences about socio-cultural characteristics that are not incorporated into the weights (bottom sections, Figure 2). The intuition here is that non-ignorable non-response within strata would make itself apparent if weighting draws the sample away from the population distribution on dimensions that are not incorporated into the weights. Weighting minimally affects the distribution of these non-weighting variables (religion, marital status, tribe) except urban bias in the Kenya sample, which is reduced. This suggests a negligible correlation between age, gender, education, and geography and unmeasured factors that are associated with the cultural variables shown in Figure 2. Consequently, the analysis provides some assurance that weighting does not increase bias for descriptive inferences. Likewise, weighting only slightly impacts public opinion estimates (see section "Comparing public opinion estimates," below) and tends to move these estimates slightly towards - rather than away from - benchmarks. Weighting does introduce bias in our estimates of experimental effects in Indonesia; we diagnose this bias in the next section. Overall, population-based weighting improves the quality of the statistics derived from our surveys, without introducing additional bias.

Replicating classic survey-experimental findings

Most social science researchers are primarily interested in public attitudes or behaviors, rather than demographic summaries. We now turn to these quantities of interest. As an initial check on the face validity of survey results derived from our samples, we use a

canonical behavioral experiment — the Tversky and Kahneman (1981) "disease problem" used to test prospect theory. Since other low-cost platforms have been shown to produce reliable replications of this and other survey-experimental findings (Berinsky, Huber, and Lenz 2012; Coppock and McClellan 2019; Mullinix et al. 2015), we include this analysis as a low bar, benchmark test with which scholars may have familiarity.

This survey experiment asks respondents to *"Imagine that your country is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed."* Then respondents are randomly assigned to see two program options which are framed either in terms of the number of lives that will be saved (the "save" condition) or the number of lives that will be lost (the "die" condition) under the two options. Within each condition, the difference between the two programs is whether the outcome is stated in certain or probabilistic terms. For instance, in the "save" condition, one program will definitively save 200 people, whereas the other program has a one-third probability of saving no people and a two-thirds probability of saving 600 people. In expectation, the payoffs of both policies are equal. The finding that has been replicated in many samples is that differences in respondents' choices should be driven primarily by their appetites for risk and that people are loss averse. Respondents tend to be more willing to take on risk to avoid losses (deaths) rather than to accrue gains (saving lives). In the experiment this manifests as larger numbers choosing the risky option in the "die" condition and the certain option in the "save" condition (Tversky and Kahneman 1981).

Table 3 shows the original results from Tversky and Kahneman (1981) among their sample of U.S. students, a replication by Berinsky, Huber, and Lenz (2012) using a US-

based MTurk sample, and results from our Facebook samples in Kenya and Indonesia.[11] Our results are largely consistent with other studies that have replicated this experiment across cultural contexts (Im and Chen 2022; Ruggeri et al. 2020). For instance, Im and Chen (2022) report that the percent of respondents who pick the certain option when framed in a positive (save) vs. negative (die) framing in South Africa is 57% and 37%, in Mexico is 59% and 37%, and in Indonesia is 51% and 33%, respectively.

These preferences manifest in the same direction in our samples, with one notable exception: the application of weights skews the result in the "die" condition in Indonesia. We explore this departure from the classic finding in Indonesia in SI Tables S2 and S3. We show that in Indonesia, the unweighted result in the "die" condition is consistent with our expectations (respondents exhibit loss aversion) (Table S2). However, the result does not hold for respondents over the age of 50 (Table S3), which suggests that the older individuals we were able to recruit are distinctive in how they respond to this experiment. Recall from the "Sampling error" section that we were unable to recruit respondents over the age of 50 from several of the strata in our sample (sampling bias). The

[11]The Indonesia data analyzed here is from the pilot version of the survey, fielded in July 2023. The SI provides more information on the pilot wave of the survey.

TABLE 3 *Replication of Tversky and Kahneman (1981) Disease Problem*

	Tversky & Kahneman		**Berinsky, Huber, & Lenz**		**Kenya Facebook sample (weighted)**		**Indonesia Facebook sample (weighted)**	
Options	Save	Die	Save	Die	Save	Die	Save	Die
Certain	72%	22%	74%	38%	61%	36%	55%	49%
Risky	28%	78%	36%	62%	39%	64%	45%	51%

Note: The table shows the proportion of respondents choosing "certain" and "risky" policies to manage a disease, when the policies are framed in terms of lives saved vs. deaths. When the policies are framed in terms of the number of lives saved, a majority of respondents prefers the certain policy. When the policies are framed in terms of the number of people who will die, the majority prefer the risky option.

resulting under-representation means that we must apply relatively heavy weights for these individuals.[12] Applying these weights to estimate experimental effects pulls the average effect towards the biased (relative to the benchmark) result among older individuals. This reflects adjustment bias. We highlight this finding because it reinforces our suggestion (introduced in the "Sampling Error" section) to monitor sampling quotas and dedicate resources to minimizing sampling bias. The finding also cautions against the uncritical use of weights to address under-representation of certain groups.

Comparing Public Opinion Estimates

We next report descriptions of public opinions and political behaviors including partisanship, political engagement, attitudes towards the president, voting behavior, and beliefs about climate change, a policy area of increasing interest to social scientists. In each case, we compare responses from Facebook samples to identical opinion questions fielded in benchmark surveys (the Afrobarometer, the LAPOP, and the Asian Barometer). Figure 4 plots the Facebook sample estimates, both weighted and unweighted, as well as benchmark estimates. In general, where there are gaps between the Facebook-derived estimates and benchmark samples, Facebook samples report greater political activity. Note that the application of weights tends to move Facebook-derived estimates towards benchmark survey estimates, reinforcing the point made in the "Adjustment error" section that adjustment bias does not appear serious. Instead, estimate accuracy tends to improve when applying weights.

Persistent gaps in Figure 4 are likely attributable to some combination of sample

[12] The average weight for respondents over 50 is 2.34.

composition and survey mode. Representation bias could account for some upward bias in estimates of political activity, since Facebook samples over-represent highly educated respondents who are likely to be more politically engaged (Verba, Schlozman, and Brady 1995). Still, survey mode could also play a role here. For example, barometer survey data might reflect under-reporting of certain political activities if social desirability bias deters respondents from reporting activities that challenge the government, such as protesting, when responding to in-person surveys. Consistent with this tendency, barometer survey respondents report lower levels of these types of activities and, in Kenya, higher levels of approval of the President. Conversely, evidence suggests that turnout is over-reported in surveys (Holbrook and Krosnick 2010), and this bias in self-reporting may be larger for in-person surveys (Jackman and Spahn 2019). Accordingly, voter turnout is the only place where Facebook samples report *lower* activity than corresponding barometer estimates. Moreover, the Facebook sample estimates are closer to the official turnout figures reported in each country (indicated by the green crosses in Figure 4). In Indonesia, Facebook estimates of the proportion of respondents who vote in most or every election are somewhat lower than both the Dynata and Asian Barometer estimates.

Consideration of political context reinforces the idea that different estimates of political activity may be in part attributable to mode effects. The Asian Barometer survey was fielded in July 2019 in Indonesia, just a few months after President Joko Widodo's re-election sparked protests and claims of election fraud from the opposition Gerindra party candidate (Suhartono and Victor 2019). In this context, Asian Barometer respondents may have been unwilling to express their true political views to an enumerator. For instance, only 1% of Asian Barometer respondents reported affiliation with the opposition Gerindra party, whose candidate received 44% of the vote in the presidential election. In the Facebook sample, many more respondents (16%) reported affiliation with the Gerindra

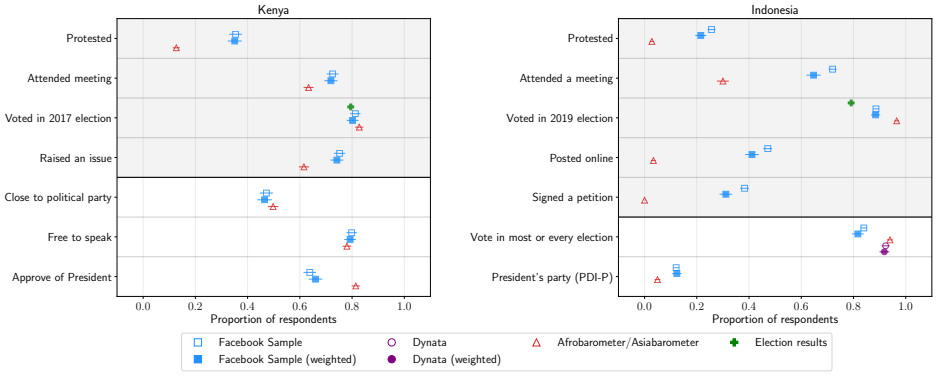


Figure 4. Political attitudes and behaviors
Note: Figure shows responses from Afrobarometer/Asian Barometer and Facebook samples in Kenya and Indonesia, to questions related to political attitudes and behaviors. It compares sample respondents according to self-reported behaviors including identifying with a political party, voting, and engaging in activities such as community meetings and protests. Estimates for both samples have been weighted using individual weights provided by Afrobarometer/Asian Barometer surveys, or raking procedure described above (for Facebook samples).

party. Similarly, more Facebook respondents reported having protested, signed a petition, posted online, or attended a meeting in the previous three years, compared with Asian Barometer respondents. Asian Barometer respondents were also more likely to report that they did not know the answer to these questions about political activity, which is curious in light of how memorable these activities can be. The high rate of "don't know" responses reported in the Asian Barometer survey suggests social pressure associated with political tensions could have led respondents to equivocate in their reporting of partisan loyalty and political participation, in the in-person survey context.

To illuminate the platform's utility for assessing policy views, we compare oenvironmental and climate change opinions in Mexico and Indonesia. In Mexico, we replicated the question wording and response options from a LAPOP question (LAPOP 2018–2019), regarding whether economic growth or environmental protection should be prioritized.

Facebook respondents were more likely than LAPOP respondents to answer that environmental protection should be given the highest priority (44%, vs. 20%), and less likely to choose responses at the economic-growth end of the Likert scale (9% vs. 18%) (Figure 5, left panel). In Indonesia, we asked multiple questions about climate change to our Facebook-recruited respondents and to respondents recruited from Dynata's online panel. Facebook respondents were less worried about climate change and less likely to understand that it is human-caused, compared with the Dynata sample (Figure 5, right panel). These differences likely stem from sample composition. The Facebook sample is more highly educated than the LAPOP sample in Mexico but less highly educated than the Dynata sample in Indonesia, and education is positively correlated with environmental concern in global surveys (Lee et al. 2015).

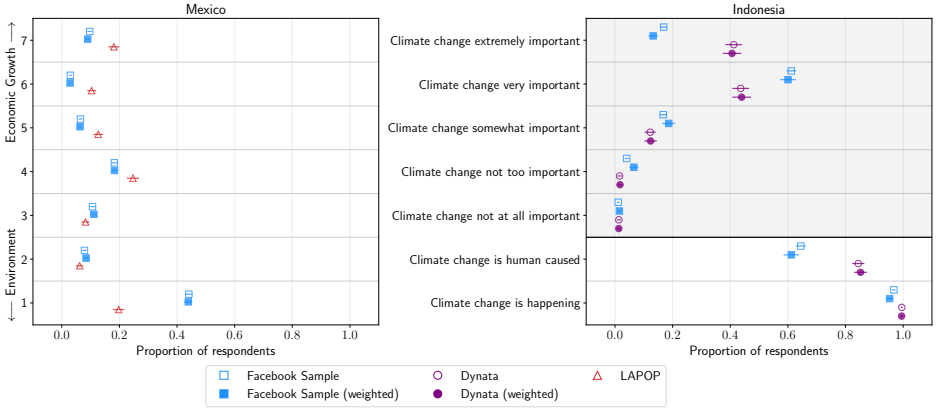


Figure 5. Public opinion estimates
Note: Figure shows responses from LAPOP and Facebook samples, to the question of whether environmental protection (1) or economic growth (7) should be given priority. Facebook results are weighted, and LAPOP results are unweighted, consistent with LAPOP documentation.

Survey costs

Researchers typically weigh survey quality concerns against the relative costs of different survey tools. For Facebook-recruited surveys, costs include platform advertisements and (optional) incentives paid to respondents. The SI provides details about the reach of our survey as advertised on Facebook, and the cost of reaching different demographic groups.[13] Facebook sampling is quite cost-effective. Not including incentives, the mean cost per completed survey was $0.16 in Mexico,[14] $0.85 in Kenya,[15] and $0.91 in Indonesia. Including incentives, the surveys cost an average of $1.03 per completed survey (ranging from $0.16 in Mexico to $1.57 in Indonesia). This is incredibly inexpensive in Mexico, and in Kenya or Indonesia the cost is comparable with the cost of recruiting online convenience samples using platforms such as MTurk or Lucid. Of course, MTurk, Lucid, or similar other platforms do not enable researchers to contact survey respondents in every country, and their user populations are much smaller than Facebook's. Realistically, in-person field surveys or online panels provide the most feasible alternative in most parts of the Global South. The costs of our Facebook samples are substantially cheaper than these alternatives. For instance, the Indonesia Dynata sample (n=1,130) cost $5.75 per completed survey. In-person surveys are even more expensive to field.

[13] These costs vary because Facebook ads are deployed using a bidding system in which hard-to-reach populations are more expensive to target.

[14] These statistics reflect the full campaign, including our low-education oversample. In the SI, we separate costs by the initial sample (without education targeting) and for the oversample

[15] Using Facebook's count of completed surveys; this falls to $0.56 if we consider all 2,323 surveys *initiated* on Qualtrics, and $0.89 if we consider all 1,452 valid surveys completed on Qualtrics used in the analysis.

Conclusion

As of 2019, Facebook was the most popular social media platform in the world with 2.3 billion users.[16] Facebook's extensive user base presents an opportunity for researchers to quickly and reliably recruit subjects from diverse countries across the world, including contexts that are underrepresented on existing online subject recruitment platforms. Our replication of the disease problem experiment suggests that, similar to other low-cost online survey platforms (Berinsky, Huber, and Lenz 2012; Mullinix et al. 2015; Coppock and McClellan 2019), Facebook is a cost-effective tool for recruiting survey-experimental subjects. The platform's availability in virtually every country makes it especially attractive in countries where other low-cost survey recruitment platforms do not reach. Our analysis provides even deeper insights into how scholars can use Facebook to recruit survey samples. Using a series of comparative analyses, we evaluate the quality of Facebook-recruited samples in three countries across the Global South, highlighting some of the advantages and shortcomings of this method and providing practical guidance to researchers on how to measure and address the platform's limitations.

We have assessed total survey error and its various components, in order to provide practical insights into where and how researchers can minimize bias in estimates derived from Facebook-recruited surveys. Coverage error favors respondents who have achieved higher levels of education, on average, than the general population. This over-representation of highly educated respondents is a structural feature of the Facebook platform in our case countries. By extension, researchers should investigate these imbalances in other target countries before deciding to use Facebook as a survey recruitment tool. Researchers can

[16] https://ourworldindata.org/rise-of-social-media.

mitigate bias associated with coverage error by targeting recruitment resources towards respondents who are under-represented on Facebook. In our case, quota sampling through the ad platform allowed recruitment of a more representative sample than if the advertising algorithm were left to its own devices. However, we do find some evidence of sampling bias in the case of Indonesia. Researchers should monitor entries into the survey from each strata and adjust their advertising campaign to ensure recruitment in all strata. Still, the back-end demographic data that Facebook uses for targeting are noisy, particularly for education and location. As a result, quota sampling cannot substitute for the use of design weights based on self-reported demographics. Researchers should also assess non-response rates across quota cells during data collection to inform an iterative sampling strategy that maximizes response rates among high-attrition groups. In general, weights reduced but did not completely eliminate demographic imbalances. The weights affected descriptive inferences only slightly, and our weighted estimates tended towards benchmark estimates, compared with unweighted estimates. We conclude that adjustment bias is minimal in our descriptive inferences. However, we do find some evidence for adjustment bias in our experimental findings, where older respondents exhibit unusual behavior and are up-weighted in the results.

Practically, the costs of using Facebook to recruit survey respondents can be quite low, though costs depend on targeting details and incentives for survey completion. In Mexico, we successfully recruited respondents without incentives. In Kenya and Indonesia, because much of the population uses rate-limited mobile internet, we provided a modest airtime credit as compensation. This incentive may have encouraged greater participation among resource-constrained respondents, but it also led to instances of gaming and viral sharing of survey links. Future research could investigate sampling strategies to leverage the social nature of the platform, for example using snowball sampling that encourages the sharing

of survey links.

Facebook suffers from certain limitations as a platform, and context matters when deciding to use Facebook to recruit subjects. The platform will be particularly successful where 1) phone and internet penetration is widespread, 2) literacy rates are high, and 3) recent census (or other benchmark) population data are available to allow weighting. Researchers should also keep in mind that users' interactions with and expectations for Facebook might influence internal validity. While we have tested this method in three competitive democracies, in authoritarian states (where Facebook is permitted) citizens might have different assumptions about the government's surveillance of their platform responses. Researchers should consider these concerns in the survey design process, and future research could consider how internal validity varies across political contexts. The popularity of Facebook is also changing over time, and the platform may not retain the large and broad user base that make it an attractive platform for survey sample recruitment. Changes to Meta's advertising policies may also change researchers' ability to target ads towards particular groups. Such changes are beyond researchers' ability to control but well within their ability to monitor in order to use the platform wisely. Moreover, our conceptual framework can be applied to other platforms that may arise to supplant Facebook in popularity.

Of course, Facebook should not supplant gold-standard, resource-intensive, in-person field surveys for recruiting nationally representative samples in the Global South. Our Facebook samples exhibit higher levels of educational achievement and a slightly different age distribution, compared with the national populations and benchmark surveys in our case countries. Correspondingly, the descriptive inferences we draw regarding political engagement and public policy views are slightly skewed towards the views of more highly educated individuals. However, our Facebook sample out-performs, at considerably

lower cost, the sample recruited from a commercial online survey firm. Facebook thus represents an opportunity to cheaply reach diverse populations. Further, as we have shown, researchers can iteratively adjust their sampling protocol, model non-response bias, and utilize weighting to mitigate sample biases. We further recommend that researchers include in their surveys at least a few benchmark comparisons similar to those we have performed here. Facebook represents a valuable tool that, when used well, could open new frontiers in public opinion research.

References

Afrobarometer. 2016. *Afrobarometer, 2016.* http://www.afrobarometer.org.

Ansolabehere, Stephen, Brian Schaffner, and Marisa Shih. 2023. *Guide to the 2022 Cooperative Election Study.* https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910/DVN/PR4L8P.

Ansolabehere, Steve, and Brian Schaffner. 2010. "Does survey mode still matter." *Political Analysis.*

Araujo, Matheus, Yelena Mejova, Ingmar Weber, and Fabricio Benevenuto. 2017. "Using Facebook Ads Audiences for Global Lifestyle Disease Surveillance: Promises and Limitations." In *Proceedings of the 2017 ACM on Web Science Conference,* 253–257. WebSci '17. Troy, New York, USA: ACM. ISBN: 978-1-4503-4896-6. https://doi.org/10.1145/3091478.3091513.

Bailey, Michael. 2024. *Polling at a crossroads.* Cambridge University Press.

Berinsky, Adam, Gregory Huber, and Gabriel Lenz. 2012. "Evaluating online labor markets for experimental research." *Political analysis* 20 (3): 351–368.

Bicalho, Clara, Melina Platas, and Leah R. Rosenzweig. 2020. *"If we move, it moves with us:" Physical Distancing in Africa during COVID-19.* Working, May. https://static1.squarespace.com/static/5410fc5ae4b0b9bdbd0cde82/t/5f0b12f8edb3507adfee3b0f/1594561273957/Social_distancing_africa_BPR2020.pdf.

Boas, Taylor C, Dino P Christenson, and David M Glick. 2020. "Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics." *Political Science Research and Methods,* 232–250.

Bond, Robert, and Solomon Messing. 2015. "Quantifying social media's political space: Estimating ideology from publicly revealed preferences on Facebook." *APSR* 109 (1): 62–78.

Broockman, David, and Donald Green. 2014. "Do online advertisements increase political candidates' name recognition or favorability? Evidence from randomized field experiments." *Political Behavior* 36 (2): 263–289.

Coppock, Alexander, and Oliver McClellan. 2019. "Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents." *Research & Politics* 6 (1): 2053168018822174.

DeBell, Matthew, Michelle Amsbary, Ted Brader, Shelley Brock, Cindy Good, Justin Kamens, Natalya Maisel, and Sarah Pinto. 2020. *Methdology Report for the ANES Time Series Study*. https://electionstudies.org/wp-content/uploads/2022/08/anes_timeseries_2020_methodology_report.pdf.

Deming, Edwards. 1944. "On errors in surveys." *American Sociological Review* 9 (4): 359–369.

Finkel, Steven E., Anja Neundorf, and Ericka Rascon Ramirez. 2023. "Can online civic education induce democratic citizenship?" *AJPS*.

Gomes, Manuel, Rosalba Radice, Jose Camarena Brenes, and Giampiero Marra. 2019. "Copula selection models for non-Gaussian outcomes that are missing not at random." *Statistics in medicine* 38 (3): 480–496.

Grewal, Sharan. 2023. "Military Repression and Restraint in Algeria." *APSR,* 1–16.

Groves, Robert, Floyd Fowler Jr, Mick Couper, James Lepkowski, Eleanor Singer, and Roger Tourangeau. 2011. *Survey methodology*. Vol. 561. John Wiley & Sons.

Groves, Robert, and Lars Lyberg. 2010. "Total survey error: Past, present, and future." *Public opinion quarterly* 74 (5): 849–879.

Grow, André, Daniela Perrotta, Emanuele Del Fava, Jorge Cimentada, Francesco Rampazzo, Sofia Gil-Clavel, and Emilio Zagheni. 2020. "Addressing public health emergencies via Facebook surveys." *Journal of Medical Internet Research* 22 (12): e20653.

GSMA. 2023. *The Mobile Economy 2023*. Technical report. https://www.gsma.com/mobileeconomy/wp-content/uploads/2023/03/270223-The-Mobile-Economy-2023.pdf.

Hartman, Erin, and Melody Huang. 2024. "Sensitivity analysis for survey weights." *Political Analysis* 32 (1): 1–16.

Hirano, Shigeo, Gabriel Lenz, Maksim Pinkovskiy, and James M Snyder Jr. 2015. "Voter learning in state primary elections." *AJPS* 59 (1): 91–108.

Holbrook, Allyson, and Jon Krosnick. 2010. "Social desirability bias in voter turnout reports." *Public Opinion Quarterly* 74 (1): 37–67.

Holliday, Derek, Tyler Reny, Alex Rossell Hayes, Aaron Rudkin, Chris Tausanovitch, and Lynn Vavreck. 2021. *Democracy Fund + UCLA Nationscape Methodology and Representativeness Assessment.* https://www.voterstudygroup.org/uploads/reports/Data/Nationscape-Methodology-Representative-Assessment-2021Dec.pdf.

Huff, Connor, and Dustin Tingley. 2015. ""Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents." *Research & Politics* 2 (3): 2053168015604648.

Im, Hohjin, and Chuansheng Chen. 2022. "To save or lose? A cross-national examination of the disease risk framing effect and the influence of collectivism." *Journal of Behavioral Decision Making* 35 (4): e2276.

INEGI. 2015. *Porcentaje de la población de 15 años y más con algún grado escolar por entidad federativa según nivel de escolaridad y sexo, años censales seleccionados 2000 a 2015.* Accessed: 20 August, 2019. http://en.www.inegi.org.mx/app/tabulados/pxweb/inicio.html?rxid=85f6c251-5765-4ec7-9e7d-9a2993a42594&db=Educacion&px=Educacion_04.

Jackman, Simon, and Bradley Spahn. 2019. "Why does the American national election study overestimate voter turnout?" *Political Analysis* 27 (2): 193–207.

Jacobson, Sophie. 2024. "Policies Can Still Create New Politics: Contemporary Causal Evidence from New York Preschool Parents." Unpublished working paper.

Jäger, Kai. 2017. "The potential of online sampling for studying political activists around the world and across time." *Political Analysis* 25 (3): 329–343.

Kapp, Julie, Colleen Peters, and Debra Oliver. 2013. "Research recruitment using Facebook advertising: big potential, big challenges." *Journal of Cancer Education* 28 (1): 134–137.

Kilavuz, Tahir, Sharan Grewal, and Robert Kubinec. 2023. "Ghosts of the Black Decade." *Journal of Peace Research* 60 (1): 9–25.

KNBS. 2010. *2019 Population and Housing Census.* Accessed: 9 June, 2020. http://www.knbs.or.ke.

Kosinski, Michal, Sandra Matz, Samuel Gosling, Vesselin Popov, and David Stillwell. 2015. "Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines." *American Psychologist* 70 (6): 543.

LAPOP. 2018–2019. *Mexico.* Accessed: 16 October, 2019. https://www.vanderbilt.edu/lapop/mexico.php.

Lee, Tien Ming, Ezra Markowitz, Peter Howe, Chia-Ying Ko, and Anthony A Leiserowitz. 2015. "Predictors of public climate change awareness and risk perception around the world." *Nature climate change* 5 (11): 1014–1020.

Litman, Leib, Jonathan Robinson, and Tzvi Abberbock. 2017. "TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences." *Behavior research methods* 49 (2): 433–442.

Lyberg, L, and H Weisberg. 2016. "Total survey error: a paradigm for survey methodology." In *The Sage handbook of survey methodology,* edited by Christof Wolf, Dominique Joye, Tom W. Smith, and Yang-chih Fu, 27–35. Sage Thousand Oaks, CA.

Manski, Charles F. 1990. "Nonparametric bounds on treatment effects." *American Economic Review* 80 (2): 319–323.

McGovern, Mark, David Canning, and Till Bärnighausen. 2018. "Accounting for non-response bias using participation incentives and survey design:" *Economics letters* 171:239–244.

Meta. 2023. *Meta reports second quarter 2023 results.* (Accessed on August 18, 2023). https://investor.fb.com/investor-news/press-release-details/2023/Meta-Reports-Second-Quarter-2023-Results/default.aspx.

Mullinix, Kevin J, Thomas J Leeper, James N Druckman, and Jeremy Freese. 2015. "The generalizability of survey experiments." *Journal of Experimental Political Science* 2 (2): 109–138.

Neundorf, Anja, and Aykut Öztürk. 2023. "How to improve representativeness and cost-effectiveness in samples recruited through meta: A comparison of advertisement tools." *Plos one* 18 (2): e0281243.

Noh, Yuree, Sharan Grewal, and Tahir Kilavuz. 2023. "Regime Support and Gender Quotas in Autocracies." *APSR,* 1–18.

Offer-Westort, Molly, Leah Rosenzweig, and Susan Athey. 2024. "Battling the coronavirus "infodemic" among social media users in Kenya and Nigeria." *Nature Human Behavior.*

Ortiz-Espina, Esteban. 2019. *The rise of social media.* Accessed: 15 August, 2024. https://ourworldindata.org/rise-of-social-media.

Pew Research Center. 2018. *Comparing Survey Sampling Strategies: Random-Digit Dial vs. Voter Files.* https://www.pewresearch.org/methods/wp-content/uploads/sites/10/2018/10/REPORT-Comparing-RDD-and-RBS-Surveys-Final-05.06.19-2.pdf.

Pötzschke, Steffen, and Michael Braun. 2017. "Migrant Sampling Using Facebook Advertisements: A Case Study of Polish Migrants in Four European Countries" [in en]. 00011, *Social Science Computer Review* 35, no. 5 (October): 633–653. ISSN: 0894-4393, accessed July 8, 2019. https://doi.org/10.1177/0894439316666262.

Ramo, Danielle, Theresa Rodriguez, Kathryn Chavez, Markus Sommer, and Judith Prochaska. 2014. "Facebook recruitment of young adult smokers for a cessation trial: methods, metrics, and lessons learned." *Internet Interventions* 1 (2): 58–64.

Rosenzweig, Yang-Yang, Leah Rand Zhou. 2021. "Team and nation: Sports, nationalism, and attitudes toward refugees." *Comparative Political Studies* 54 (12): 2123–2154.

Rotondi, Valentina, Ridhi Kashyap, Luca Maria Pesando, Simone Spinelli, and Francesco C Billari. 2020. "Leveraging mobile phones to attain sustainable development." *Proceedings of the National Academy of Sciences.*

Ruggeri, Kai, Sonia Alí, Mari Louise Berge, Giulia Bertoldo, Ludvig D Bjørndal, Anna Cortijos-Bernabeu, Clair Davison, Emir Demić, Celia Esteban-Serna, Maja Friedemann, et al. 2020. "Replicating patterns of prospect theory for decision under risk." *Nature human behaviour* 4 (6): 622–633.

Ryan, Timothy. 2012. "What makes us click? Demonstrating incentives for angry discourse with digital-age field experiments." *The Journal of Politics* 74 (4): 1138–1152.

Samuels, David, and Cesar Zucco Jr. 2014. "The power of partisanship in Brazil." *AJPS* 58 (1): 212–225.

Santoso, Lie Philip, Robert Stein, and Randy Stevenson. 2016. "Survey experiments with Google consumer surveys." *Political Analysis* 24 (3): 356–373.

Ševčíková, Hana. 2020. *Explorer of World Population Prospects.*

Suhartono, Muktita, and Daniel Victor. 2019. "Violence erupts in Indonesia's capital in wake of presidential election returns." *New York Times* (June).

Sun, BaoLuo, Lan Liu, Wang Miao, Kathleen Wirth, James Robins, and Eric J Tchetgen Tchetgen. 2018. "Semiparametric estimation with data missing not at random using an instrumental variable." *Statistica Sinica* 28 (4): 1965.

Ternovski, John, and Lilla Orr. 2022. "A note on increases in inattentive online survey-takers since 2020." *Journal of Quantitative Description: Digital Media* 2.

Tversky, Amos, and Daniel Kahneman. 1981. "The framing of decisions and the psychology of choice." *Science* 211 (4481): 453–458.

United Nations. 2019. *World Population Prospects Volume 1, Volume 1,* ISBN: 978-92-1-148327-7.

Verba, Sidney, Kay Schlozman, and Henry Brady. 1995. *Voice and equality: Civic voluntarism in American politics.* Harvard University Press.

World Bank. 2020. *World Bank Open Data.* Accessed May 20, 2020. https://data.worldbank.org/indicator.

Zhang, Baobao, Matto Mildenberger, Peter Howe, Jennifer Marlon, Seth A Rosenthal, and Anthony Leiserowitz. 2020. "Quota sampling using Facebook advertisements." *Political Science Research and Methods,* 558–564.

Survey sampling in the Global South using Facebook advertisements: Online Appendix

Table of Contents

Sampling error **S-1**

Recruitment and sampling **S-5**
Mexico: sampling by administrative unit S-8
Indonesia: Sampling by administrative unit S-10
Kenya: sampling by geolocation S-12

Prospect Theory Experiment in Indonesia **S-16**

Survey Costs **S-18**
Variation in costs by strata S-18
Costs by round of data collection S-20
Variation in costs for specific populations of interest S-22

Weighting **S-23**

Sampling error

As mentioned in the main text, it is useful to examine the composition of the sample of ad clickers, so that researchers can adjust their quota sampling to target groups that are under-represented. Figure S1 shows the sample of ad clickers and the Facebook population in each country, based on the Facebook-inferred demographics used for targeting. The Facebook population serves not as a benchmark here but rather as an illustrative basis

for comparison, since the goal is not to reflect the Facebook population but instead the national population in each country.

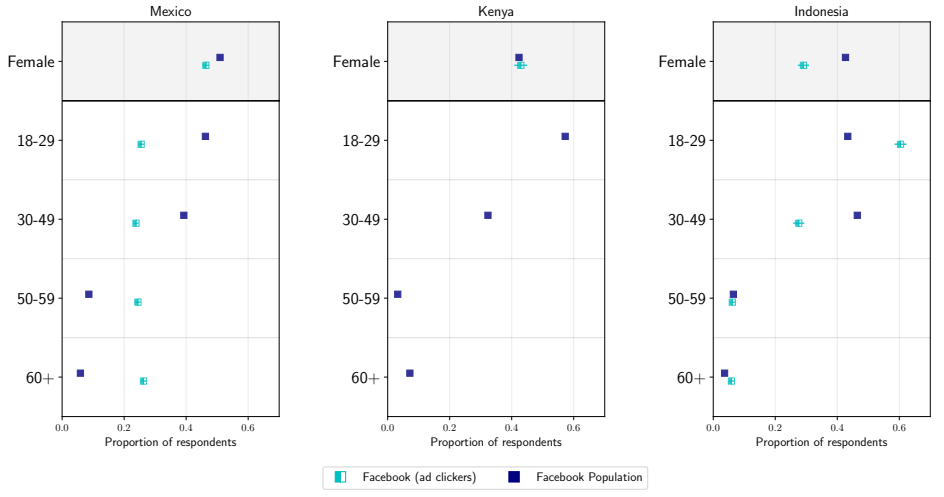


Figure S1. **Sampling error:** *The figure shows the proportion of the Facebook population (dark, solid squares) and ad clickers (light, partially filled squares) from each age and gender group used in our sampling quotas. Demographics are based on Facebook's back-end data about users. Note that in Indonesia ads could only be targeted at respondents over 21 years of age so the "18-29" year category was actually "21-29" for the Facebook ad clickers.*

As reported in the main text, we failed to fill some strata. Here, we further discuss our ability to fill our ad quota cells and the match between Facebook-reported and self-reported characteristics. In Mexico, we find a good distribution of responses. In each of our 128 geographic-demographic target cells we collected responses from between 17 and 77 individuals, with a median of 41 individuals.[1] In Kenya, we were less successful in filling our 66 strata. The quota targets for each stratum were set according to the Afrobarometer weights associated with each stratum's geolocation, which are based on population.[2]

[1] 1st quartile: 33, 3rd quartile: 46.
[2] For the province-level oversampling of older and less-educated respondents, we set a

Therefore, the number of respondents targeted per stratum ranged from 4 - 167. Ultimately, we filled 17 of our 66 target strata; the strata that fell short were missing a median of three respondents.[3] In Indonesia, we recruited at least one individual from 200 of the 272 strata that we targeted.

Two main factors contributed to the failure to fill some strata. First, we manually closed several of our survey strata because the advertising cost per completed survey was too high ($5 or more per respondent). Second, because of concern about viral sharing and completions of surveys that were not recorded by Facebook, we typically ended survey ads slightly before the corresponding stratum was filled.

In the Mexico and Indonesia samples, 87% and 77% of respondents, respectively, reported ages consistent with their Facebook advertisement strata.[4] There were no systematic patterns in which age categories were prone to mismatches. As reported in the main text, when we targeted ads to Kenyans 32 years and older, on the other hand, only 47% of respondents who reached our survey from these ads were indeed 32 years old or older. The ages of these respondents ranged from 19 to 48 years old, with a mean of 31 years.

Facebook's gender data were more accurate in Mexico and Kenya, but slightly less so in Indonesia. In Mexico, Facebook assigned a gender that matched respondents' self-reported gender for 99% of the respondents. In Kenya, Facebook ads performed almost as well, with 90% of respondents reporting a gender identity in the survey that matched the ad that targeted them. The 10% of respondents who were recruited from an ad that was targeted

quota of 5 respondents per stratum.

[3] Min: 1, max: 52.

[4] We asked age through an open-ended question. Some respondents (n=52) did not provide their age, and these individuals are not included in this calculation.

toward the opposite gender might have resulted from respondents sharing ads with friends so that they could also benefit from taking the survey and receiving 50 KES in airtime. It would not be surprising that greater sharing would have occurred in the context of the incentivized survey in Kenya, compared to the non-incentivized survey in Mexico.

Targeting by education was precise in Mexico but not in Kenya. In our second round of data collection in Mexico we targeted respondents with a high school degree and lower levels of education. We cannot tell exactly how Facebook maps these categories from the US education system onto the Mexican education system. Most crucially, the Mexican system offers technical school degrees that students complete around the same ages as American students graduate from high school. If we consider individuals with these degrees to be "high school graduates," then Facebook correctly categorized 70% of the respondents in the low-education group. In Kenya we targeted Facebook users who did not specify their educational attainment in an attempt to recruit less educated Kenyans.[5] The 31 respondents recruited by these ads did have slightly lower levels of education, on average, than the rest of our sample.[6] However, this was not an exact targeting strategy: only 13% of the sample recruited from ads targeting those with an unspecified education level self-reported having some secondary school or less, compared to 3% of respondents recruited from all other ads.

Geographic targeting attained similar accuracy levels in Mexico and Kenya but was less accurate in Indonesia. In Mexico, 67% of respondents reported living in a municipality

[5]Among the population of Facebook users 18 years and older in Kenya, Facebook reports an "unspecified" education level for 40% of people.

[6]For example, 10% reported that primary school, informal school, or no school was the highest level of education they had attained, compared with 2% in the rest of our sample. Only 23% of this subsample reported completing university or post-graduate education, relative to 34% in the rest of our sample.

that matched their Facebook advertising target cell. Most of these errors were a function of mismatches within (rather than between) the four main regions of Mexico: 92% of respondents had matching self-reported and Facebook-advertised regions. Most Kenyan respondents were targeted geographically via clusters of Afrobarometer coordinates, which could fall across multiple provinces. Therefore, we first checked whether the province corresponding to the respondent's self-reported location matched the province of at least one of the corresponding Afrobarometer coordinates. Using this definition, we achieved a 73% match rate between the Facebook target province and the Afrobarometer province.

Recruitment and sampling

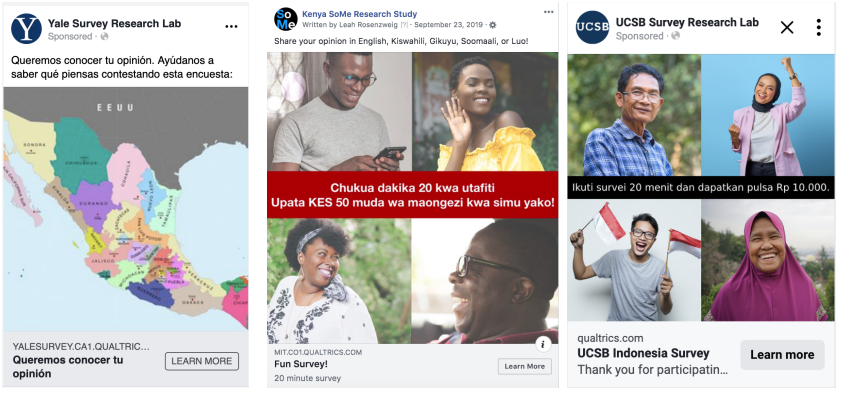
To recruit respondents for our surveys, we first created Facebook pages representing our survey campaigns, and placed ads from these pages to target people living in Mexico and Kenya. An example of these ads is shown in Figure S2. After clicking on the Facebook ad, respondents were sent to a survey hosted on Qualtrics. For Kenyan respondents, the first survey question asked respondents to choose from one of five possible languages (English, Kiswahili, Kikuyu, Luo, and Somali) in which to take the survey. For Mexican respondents the survey was administered in Spanish. Upon completing the survey, respondents were directed to a thank you page with an embedded Facebook "Pixel" which allowed Facebook to track which of the users that clicked on the ad actually completed the survey.

To help advertisers maximize their budgets, Facebook attempts to optimize ad placement according to a specific campaign objective specified by the advertiser. In our case, we used "conversion" targeting to optimize for survey completions as measured by these pixels. The Facebook targeting algorithms may introduce selection bias. To address this bias, we needed to develop a strategy to broaden the diversity and representativeness of the sample.

We use quota sampling approaches modeled on well-respected, in-person representative samples drawn from our case countries. Facebook allows advertisers to define "custom audiences" in order to target ads based on a number of personal characteristics. There are some constraints on how scholars can target their surveys since Facebook does not encode and make available all possible strata. Additionally, depending on the nature of the ad, it may be considered discriminatory to target groups based on observable characteristics such as race or gender. For example, ads involving housing, employment, or credit are subject to a limited set of targeting options.[7] Despite these constraints, we were able to design target cells based on gender and age, which are the demographic characteristics used in benchmark nationally representative surveys in these countries. We also targeted respondents by geography, again in a manner that was modeled on the geographic stratification used by in-person surveys.

Facebook allows for two different types of geographic targeting strategies. First, researchers can directly target audiences by providing a latitude and longitude of interest, as well as a radius defining the catchment area. Second, researchers can use Facebook's predefined geographic entities, which it classifies as large/medium/small "geo-areas", metro areas, cities/subcities, and neighborhoods/subneighborhoods. These entities generally correspond to known administrative units, but the correspondence is not perfect. For example, in Mexico, municipalities were alternately classified as subcities or medium geo-areas. In Kenya, targeting was available for each of the country's eight provinces, but not for all 47 counties, which have been the predominant unit of administrative organization since devolution in 2010. The availability of county targeting does not seem to be an issue

[7]For details, see: https://developers.facebook.com/docs/marketing-api/audiences/special-ad-category



(a) Mexico *(b) Kenya* *(c) Indonesia*

Figure S2. Example Facebook advertisement inviting respondents in one of our quota cells to participate in our public opinion survey.

of granularity, since it *is* possible to target specific areas of Nairobi such as Mathare (an informal settlement) or Kilimani Estate (an upscale neighborhood).

In Mexico, we target respondents according to Facebook's defined geographic entities. In Kenya, we target primarily based on latitude and longitude, although we supplement this sample with respondents from hard-to-reach groups recruited at the province level. We are therefore able to examine the viability of both these approaches to geographic targeting.

Without using quota-based ad cells, we expect that our sample would be composed of individuals who are very similar to each other. Since we target strategically, our sample includes respondents from throughout the age, gender, and geographic distributions of our case countries' residents. Below, we describe our sampling approaches in more detail.

Mexico: sampling by administrative unit

To draw a nationally representative sample of Mexican residents, we targeted Facebook users by age, gender, and geographic location. In order to gauge baseline interest in taking the survey without compensation, we did not pay respondents for their participation in the short survey. Our geographic sampling protocol mirrors the procedure used by Latin American Public Opinion Project (2017) (LAPOP) to sample within small, medium, and large municipalities.[8] However, rather than using arbitrary population size cutoffs to categorize Mexico's 2,456 municipalities[9], we group municipalities according to the distribution of Mexican residents across them. Specifically, using census data we calculated that one quarter of Mexicans reside in municipalities with fewer than 53,442 residents; one quarter in municipalities with 53,443–220,292 residents; one quarter in municipalities with 220,293–661,176 residents; and one quarter in municipalities with more than 661,177 residents. Therefore, we assign each municipality to one of these four categories. Within each of the four major regions of the country, we target these quartiles with equal weights, to gather approximately equal numbers of residents from each quartile in each region. To ensure a representative sample within the cells defined by region and size of municipality, we further stratified based on gender (male, female) and age (18-29; 30-49; 50-59; 60+) categories. These age categories are based on the age groups reported by the Mexican census bureau's municipal-level population summaries (INEGI 2015).

In total, this sampling procedure created 128 cells (4 regions $\times$ 4 municipality size

[8]LAPOP collects a sample that is stratified by four geographical regions, the size of municipality (100,000+ inhabitants; 25,000-100,000 inhabitants, and less than 25,000 inhabitants), and urban and rural areas within municipalities.

[9]These municipalities include the 16 municipal jurisdictions within Mexico City, which is designated as a unique state-level jurisdiction and divided into distinct municipalities.

groups × 2 genders × 4 age groups).[10] We collected a minimum of 25 responses from each cell, turning our ads off at regular intervals for cells that had been filled. We first collected 1,113 responses on 17 August 2019 by turning on ads for quotas located in Central Mexico. We then advertised to cells in the rest of the country on 18 August 2019, collecting 3,239 responses. After dropping individuals with self-reported ages under 18 (n=165) and one individual without a recorded stratum[11], we were left with a final sample size of 4,396. Respondents were not compensated for their participation in the survey, and we did not collect any identifiable data from any individual respondent.

After this initial data collection, we found that our sample underrepresented Mexicans whose highest level of education completed was high school or less. To correct this imbalance, we conducted a second round of data collection in which we targeted respondents in each of our previously constructed geographic-demographic strata who had no more than a high school education. After collecting this low-education sample and dropping individuals with self-reported ages under 18 (n=165);[12] those who did not report their age, gender, geographic location, or education level; and one individual without a recorded stratum,[13] our sample contained 5,168 individuals.

[10]One limitation of the Facebook advertising platform is that custom audiences cannot target more than 250 unique geographic locations, which means that a single ad could not be used to target a quartile with more than 250 municipalities. Because a number of our low-population cells contained more than 250 Mexican municipalities, we split low-population, high-municipality cells into multiple "sub-cells". We thus ran Facebook advertisements on 184 unique strata. However, we pool these "sub-cells" in our analysis so that all respondents are assigned to one of the 128 core quota cells.

[11]This likely occurred because the user inadvertently modified the Facebook Ad url which contained quota-related embedded data.

[12]Our ads only targeted individuals older than 18, per the IRB approval for the project. Younger individuals could have entered the sample if Facebook had inaccurate information about their age. We dropped these respondents to ensure compliance with our human-subjects research approval.

[13]This likely occurred because the user inadvertently modified the Facebook Ad url

Indonesia: Sampling by administrative unit

In Indonesia we targeted users by geography, gender, and age. Specifically, we placed ads targeting 34 Indonesian provinces,[14] following the provincial targeting of the 2019 Asian Barometer. Within each province, ads were targeted to men and women of four different age categories (21-29,[15] 30-49, 50-59, 60+). In total, this sampling procedure created 272 cells (34 regions $\times$ 2 genders $\times$ 4 age groups).

We collected a minimum of 10 responses from each cell, turning our ads off at regular intervals for cells that had been filled. We offered respondents Rp. 10,000 ($\sim$ \$0.65) in airtime on local telephone carriers as compensation for taking the survey. This incentive was available to respondents who provided a mobile phone number from one of the following carriers: Axis Indonesia, Indosat Indonesia, SmartFren Indonesia, Telkomsel Indonesia, Tri Indonesia, or XL Indonesia. After removing respondents who did not complete the survey,[16] the final sample included 3,277 individuals.

In Indonesia, the data for the Prospect Theory experiment were collected during a pilot wave of the survey, fielded from July 5 through July 18, 2023. The pilot sample contained 2,294 individuals. The experiment was not included on the wave of the survey from which our main findings are reported.

which contained quota-related data.

[14] In 2022, 4 new provinces were split from previously existing provinces, so that there were actually 38 provinces when we fielded the survey. However, the Facebook interface had not been updated to reflect this disaggregation and instead tagged users within the new provinces as residing within the larger, previously existing provinces.

[15] These age categories are kept consistent with the ranges in Mexico, with the exception that the youngest age group begins at 21 instead of 18 due to Facebook restrictions on targeting to teenage populations, which begin at 21 in Indonesia.

[16] We required individuals to report their age and gender and, thus, did not need to remove those who failed to do so.

TABLE S1 *Classification scheme of Mexican states into regions for the purpose of quota sampling*

State	**Region**
Ciudad de México	Centro
Hidalgo	Centro
México	Centro
Morelos	Centro
Puebla	Centro
Querétaro de Arteaga	Centro
Tlaxcala	Centro
Aguascalientes	Centro Occidente
Colima	Centro Occidente
Guanajuato	Centro Occidente
Jalisco	Centro Occidente
Michoacán de Ocampo	Centro Occidente
Nayarit	Centro Occidente
Baja California	Norte
Baja California Sur	Norte
Chihuahua	Norte
Coahuila de Zaragoza	Norte
Durango	Norte
Nuevo León	Norte
San Luis Potosí	Norte
Sinaloa	Norte
Sonora	Norte
Tamaulipas	Norte
Zacatecas	Norte
Campeche	Sur
Chiapas	Sur
Guerrero	Sur
Oaxaca	Sur
Quintana Roo	Sur
Veracruz de Ignacio de la Llave	Sur
Tabasco	Sur
Yucatán	Sur

Kenya: sampling by geolocation

In Kenya, we targeted ads according to gender (male, female) and geography. Respondents were compensated with 50 Kenyan Shillings' (~ $0.50) worth of airtime sent to their phones.[17] The Facebook ad clearly stated that this was the incentive for participation. We used a geographic quota-based approach to mimic the Afrobarometer sampling strategy. We first obtained a list of the 227 site locations from the 2016 Afrobarometer. We queried Facebook to obtain the number of users within 20km of these clusters, and dropped any site that was associated with no daily users or less than 1,000 monthly active users, for one or both genders (n=63). Because many of the remaining sites were close to each other in population-dense areas of the country, and because we were concerned that it would be hard to include 164 different locations in the sampling plan, we clustered these sites into 25 different groups and calculated the centroid of each group. Then, we targeted audiences within 12 miles (~ 20 km) of these centroids. Within these clusters, we stratified respondents by gender. More details on the approach can be found in Figures S3 and S4.

Anticipating that we would have a hard time reaching less educated and older respondents, we also created two ads in each of the eight provinces[18] of Kenya to target users 32 years and older, and users with an "unspecified" education level.[19] For these 16

[17] "Airtime" is mobile credit that can be used for calling or data.

[18] Kenya no longer uses provinces as the country's primary geographic unit, which is now the county. However, provinces were the administrative units available on the Facebook targeting interface.

[19] Since Facebook's education levels mimic the U.S. system, the available targeting levels do not correspond to those in Kenya. We guessed that those who did not finish primary school (or who otherwise had little schooling) might have left their education blank or Facebook would be unable to impute their education level, and therefore would be assigned to this category. Our understanding from conversations with a Facebook marketing advisor is that Facebook estimates education using a range of data sources including a user's location, websites visited, pages liked, and information posted on the

ads targeting the province level, we excluded each province's capital in an attempt to reach more rural respondents. We set a quota target of five respondents for each of these 16 ads.

After removing respondents who did not complete or consent to the survey, did not report their age or gender, or finished the survey in less than five minutes, as well as removing duplicate entries, our sample contained 1,528 respondents.

Figure S3. K-means algorithm for obtaining Afrobarometer cluster targets

1. Import the list of 227 AfroBarometer survey coordinates.
2. Query the Facebook marketing API to get audience estimates of the number of people within 20 kilometers of each coordinate. Drop coordinates with less than 1,000 male and/or female monthly active users, and coordinates with no male and/or female daily active users (n = 63).
3. Apply k-means clustering to the remaining 164 coordinates, with a pre-specified number of 25 clusters. Since the results of k-means depend on the random initialization, we experimented with random seeds until we achieved an allocation for which none of the 20-km radii around the chosen centroids overlapped. This ensured that each centroid could be used to target a distinct audience (using a radius of 12 miles).
4. Determine the total count of 2016 respondents associated with each of the 25 centroids. This count of respondents was used to determine the total weight of the centroid coordinates, which in turn dictated the number of respondents we looked for in the stratum.

user's profile page.

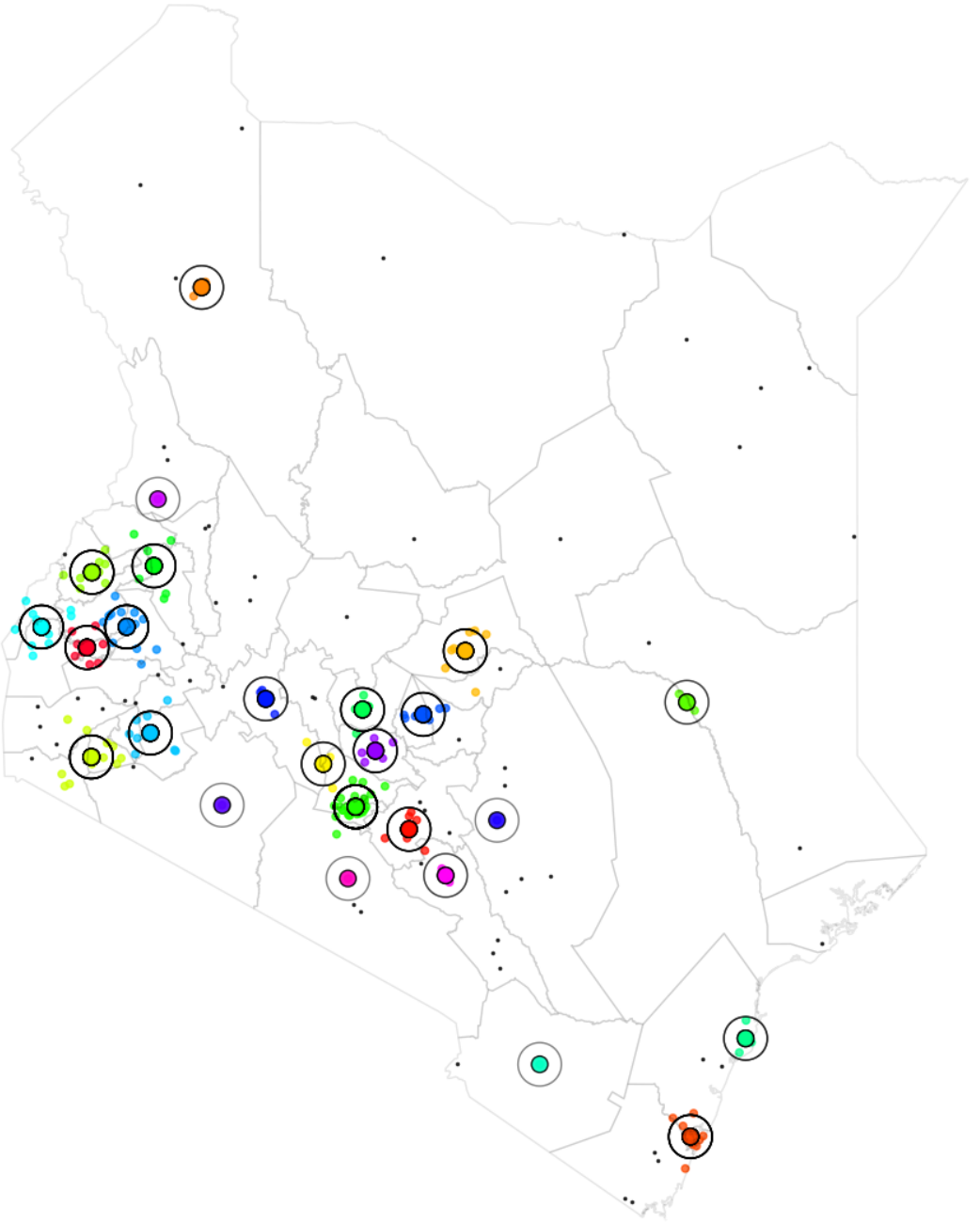


Figure S4. This map shows how Afrobarometer clusters were grouped together using the K-means algorithm. The colored dots with black outlines represent the 25 final K-means centroids used to target Facebook ads in the full study. The black circles represent 20km radii about these centroids. The colored dots with no outline represent the Afrobarometer survey clusters corresponding to each K-means centroid. The small black dots represent Afrobarometer clusters that were dropped because they had too few Facebook users.

Attention Checks in Kenyan Survey. Following previous work on identifying "shirkers," in Kenya we included two questions in the survey to check whether respondents were paying attention and answering honestly (Berinsky, Margolis, and Sances 2014; Berinsky et al. 2019). First, more than 20 questions after respondents were originally asked their age, we again said, *"To confirm you are paying attention to the survey questions, please tell us again what is your age?"* If respondents were clicking through and entering nonsense responses, we would not expect them to recall the false age they quickly entered the first time. 98% of respondents who answered both questions entered the same age in both fields and 2% entered different ages. In addition, 7% of respondents failed to answer both of these questions. Specifically, 4% did not report their age when asked the first time and 3% did not answer the second age question, which could be because they did not want to be caught falsifying responses or clicking through without paying attention.

We also included a second, more sophisticated attention check question. This question initially read like a real question, but then asked respondents to follow a set of arbitrary instructions instead of actually answering the question. The question read:

> *We are interested in how people conceive of democracy. We also want to know if people pay attention to survey questions. To show you are paying attention, please answer the question below and put the letter 'k' in the blank space next to the 'other' response. That's right, please select your real answers and put 'k' in the 'other' response. What if anything does "democracy" mean to you? (please check all that apply)*

Interestingly, only 54% of respondents passed this attention check, suggesting that this might have been a particularly hard attention check. Some respondents, especially those with low levels of education, may have found it confusing or illogical; indeed, there is a

statistically distinguishable difference in level of education achieved between those who passed this attention check question and those who did not. Based on these findings, we suggest that researchers include multiple attention checks of varying difficulty (Berinsky et al. 2019), to ensure that the questions are testing attention and honesty rather than respondent comprehension or sophistication.

Prospect Theory Experiment in Indonesia

In the main text, we present weighted results from our replication of Tversky and Kahneman (1981) classic experiment showing loss aversion in people's appetites for risk. When we apply weighting to the analysis for the Indonesia sample, this result does not hold, as shown in the final column of Table 3. Here, we probe the reason for this discrepancy with classic experimental findings and show that the finding results from adjustment bias in our pilot survey.

First, we examine the unweighted results from the experiment. Table S2 shows the weighted and unweighted result from the sample, and shows that, without applying weights, the result is consistent with the classic experimental findings. Respondents prefer the risky option in the "die" condition and the certain option in the "save" condition. However, this finding of loss-aversion disappears when we apply weights.

Next, we explain the null weighted result in the "die" condition. To do so, we examine results within sub-groups of respondents defined by age, education, and gender, which are the variables we use to construct the weights. If experimental results are correlated with the weighting variables, this would account for the discrepancy between our expectations and weighted findings. Table S3 shows experimental results by age, gender, and education groups. Readers will note that results for each group are as we expect, with the exception

TABLE S2 *Weighted and Unweighted Replication of Tversky and Kahneman (1981) Disease Problem, in Indonesia*

	Indonesia (unweighted)		**Indonesia (weighted)**	
Options	Save	Die	Save	Die
Certain	53%	45%	55%	49%
Risky	46%	55%	45%	51%

Note: The table shows the proportion of respondents choosing "certain" and "risky" policies to manage a disease, when the policies are framed in terms of lives saved vs. deaths. The first two columns show the results in Indonesia without applying weights, whereas the second two columns show the results when the results are weighted. The expectation is that when the policies are framed in terms of the number of lives saved, a majority of respondents prefers the certain policy. When the policies are framed in terms of the number of people who will die, the majority prefer the risky option.

of respondents over 50, who exhibit equal probabilities of choosing the two programs. This is crucial, since this age group is under-represented in our sample and, thus, up-weighted quite heavily. These respondents have a mean weight of 2.34. This result suggests that the failure to replicate the classic finding in our weighted sample, within the "die" condition, is attributable to heavily weighting older respondents who seem to respond to the experiment in an unusual way.

TABLE S3 *Replication of Tversky and Kahneman (1981) Disease Problem in Indonesia, by Age, Education, and Gender, for Respondents Assigned to the "Die" Condition*

	Education		**Gender**		**Age**	
	College	**No college**	**Female**	**Male**	**Under 50**	**Over 50**
Certain	45%	46%	44%	45%	46%	50%
Risky	55%	54%	56%	55%	54%	50%
N	1,241	1,053	1,105	1,189	2,135	159

Note: The table shows the proportion of respondents choosing "certain" and "risky" policies to manage a disease, when the policies are framed in terms of lives lost. Results are unweighted, and sample sizes for each group are shown in the bottom row.

Survey Costs

A key question for researchers is whether Facebook surveys are cost-effective. There are two types of marginal costs when using Facebook to field surveys: (1) advertising costs and (2) incentive costs associated with paying respondents (as we did in Kenya and Indonesia). Here, we focus on advertising costs, which we quantify in terms of cost per click and cost per completed survey. Table S4 shows the reach and total spending of our survey campaigns.

TABLE S4 *Reach of survey*

	Mexico	**Kenya**	**Indonesia**
Impressions	492,564	649,264	7,153,465
Reach	439,056	318,960	4,241,057
Clicks	24,314	8,206	23,815
Survey results	5,167	1,528	3,277
Total spent	$847.76	$1,293.91	$2,992.38
Click through rate (%)	0.049	0.013	0.003
Completion rate (%)	0.010	0.002	0.0005
Cost per click ($)	$0.035	$0.16	$0.13
Cost per completed survey ($)	$0.16	$0.85	$0.91

Note: Impressions measure how often the ads were on screen for the first time to the target audience. Reach refers to the number of people that saw the ad at least once. Clicks measures the number of people who clicked on the ad to enter the survey. Survey results reflect the number of high-quality and complete responses to the Qualtrics survey. The click through rate reflects the number of clicks divided by the number of impressions, and completion rate is the number of completed surveys divided by the number of impressions. Costs include only Facebook advertising costs (listed under total spent); they do not include survey incentives.

Variation in costs by strata

Facebook ads are deployed using a bidding system in which hard-to-reach populations are more expensive for advertisers to target. One feature of this bidding-based advertising system is that respondents in some strata costs more to recruit than those in others. This

may result from the fact that many advertisers are competing for these respondents (driving up the price in the market), or that there are fewer respondents of the target type available. Other factors driving costs could plausibly include the timing of the ads (for example, ads might be more costly to place near the holiday shopping season or political elections), and the attractiveness of the ads themselves (for example, interesting ads might generate a higher click through rate). Short of actually placing survey ads, it is hard to obtain information on targeting costs, so in Table S5, we provide evidence of this variation from our own data collection.

We find that click through rates can vary dramatically across strata, from 2-13% in Mexico, 0.7-6% in Kenya, and 0.11-1.2% in Indonesia. Similarly, there is a large spread in survey completion rates. These differential response rates are translated into different costs per click and ultimately, to dramatically different costs per completed survey. In Mexico, we paid $0.07 - 0.55 per survey, but in Kenya and Indonesia the spread was much larger, from $0.05 - 22.07 in Kenya (despite the fact that we periodically switched off high-cost strata) and from $1.99 - $16.22 in Indonesia.

Generally, we found that targeting in Mexico was cheaper and more successful than in Kenya or in Indonesia. This could be the case for two reasons: (1) the higher levels of Facebook usage in Mexico, and (2) the municipalities we used for geographic targeting in Mexico had more Facebook users in them than the smaller cells we defined using geolocated centroids and radii in Kenya. Some provinces in Indonesia are sparsely populated as well. Although we incentivized respondents in Kenya and Indonesia (and the cost for this is not included in the tables presented here), any potential increase in response rates due to the use of this incentive does not appear to have been sufficient to compensate for country-level differences.

TABLE S5 *Cost of Facebook targeting*

	Mexico			Kenya			Indonesia		
	Median	Max	Min	Median	Max	Min	Median	Max	Min
Click through rate (%)	5.56	13.25	2.16	1.2	6.0	0.7	0.41	1.2	0.11
Completion rate (%)	1.24	4.29	0.29	0.2	2.8	0.0	0.02	1.84	0.002
Cost per click ($)	0.033	0.066	0.018	0.15	0.47	0.02	0.12	0.77	0.025
Cost per completed survey ($)	0.13	0.55	0.069	1.15	22.07	0.05	1.99	16.22	0.019

Note: The click through rate and the completion rate are defined with respect to the number of impressions. The mean, maximum, and minimum are computed over quota targeting cells. Note that cost per completion does not include incentives and, for Mexico and Kenya, is calculated with reference to Facebook-reported completed surveys. For Indonesia, the completion rate and cost per complete are calculated using those who completed the survey as reported by Qualtrics rather than Facebook, because the pixel did not fire reliably in this case.

Costs by round of data collection

We next show variation in costs across types of respondents. In Mexico, we completed two rounds of data collection. The first round did not target respondents by education, but we found that the sample was skewed towards higher-education respondents. Thus, we conducted a second round of data collection in which we layered education on top of the geography × gender × age cells that we targeted in our initial data collection. For each of our initial sampling cells, we collected responses from individuals whom Facebook identified as having no more than a high school degree. Table 4 of the main text presents the reach and cost of data collection based on the full sample – i.e., including this second round of data collection targeting low-education respondents. In total, our survey reached 439,056 respondents, and we obtained 24,314 link clicks and 5,313 survey takers. Overall, the total cost per complete was $0.16. In Tables S6 and S7, we present the same statistics separately for the initial sample and the low-education oversample. The table shows the higher cost of surveying hard-to-reach respondents such as those with a high school or lower level of education.

In Kenya, we completed all data collection simultaneously, but included supplementary

clusters to oversample older people and people with unspecified education levels. Our initial k-means targeting strategy was geographically restrictive (aiming for circles with a 20km radius), whereas our oversampling strategy targeted at the province level. In other words, our oversample included harder-to-reach respondents, but we allowed Facebook to search for these respondents over a more permissive area. As a result, strata in our initial sample actually saw *higher* median and maximum costs to recruit, although the cheapest respondents in our initial sample (presumably in dense areas that were easy to target) were cheaper than the cheapest respondents in our oversample.

TABLE S6 *Reach of surveys, by education*

	Mexico		Kenya	
	Initial	Oversample	Initial	Oversample
Reach	251,742	187,314	282,190	36,770
Impressions	273,649	218,915	599,388	49,876
Clicks	16,265	8,049	7583	623
Survey results	4,380	933	1,145	66
Total spent	$354.03	$493.73	$1,216	$77.91

Note: Reach of surveys in Mexico and Kenya, split into main sample and low-education oversamples. Survey results are defined by Facebook here. Total spent does not not include incentives.

TABLE S7 *Cost of Facebook targeting in Mexico and Kenya with and without low-education oversample.*

	Mexico						Kenya					
	Initial			Oversample			Initial			Oversample		
	Med	Max	Min	Med	Max	Min	Med	Max	Min	Med	Max	Min
Click through rate (%)	6.65	21.29	2.59	4.38	10.11	1.11	1.17	6.05	0.73	1.29	3.10	0.84
Completion rate (%)	1.97	7.59	0.40	0.42	2.20	0.06	0.15	2.77	0.01	0.16	0.46	0.02
Cost per click ($)	0.019	0.049	0.01	0.07	0.18	0.024	0.18	0.47	0.02	0.09	0.33	0.04
Cost per completed survey ($)	0.07	0.31	0.024	0.58	4.50	0.08	1.19	22.07	0.05	0.78	8.72	0.22

Note: The click through rate and the completion rate are defined with respect to the number of impressions. Costs shown do not include incentives.

Variation in costs for specific populations of interest

In Table S8, we present the average cost of targeting different respondent types. The estimates shown in the table reflect the total ad spend for ads targeting each of these types of people, divided by the number of respondents who completed the survey in that category (as self-reported in the survey data), for each country. For instance, the cost per complete for women is the sum of ad spending for the ads targeting women across all geographic areas, divided by the number of complete survey responses from individuals who self-identify as women. Using this approach, we estimate that Kenyan women are over 4 times more costly to target than Kenyan men, whereas rural respondents are over 10 times more costly to target than urban respondents on average, in Kenya. We do not find such high discrepancies in the cost of contacting men and women in Mexico, and the difference in cost for surveying respondents from small and large municipalities are not as starkly different as in Kenya. Table S8 indicates that survey costs can vary dramatically by strata and by the nature of a given survey round. This variation introduces practical tradeoffs and might create disincentives to target certain populations, and these disincentives can be substantial.

TABLE S8 *Cost per response from different respondent types*

	Mexico FB	Kenya FB	Indonesia FB
Overall	$0.16	$0.85	$0.91
Men	$0.14	$0.33	$1.13
Women	$0.19	$1.50	$0.71
Urban	$0.13	$0.22	–
Rural	$0.19	$2.61	–

Note: This table sums the total spend for ads specifically targeting the specified group and then divides by the numbers of respondents in the category, as self-reported in the survey data. Afrobarometer sites are classified as rural, urban, or both. When aggregating to k-means clusters, we defined "rural" clusters for which there were more rural than urban Afrobarometer sites (n=17), and "urban" k-means clusters as those with more urban or an equal number of urban and rural (n=8). In the Mexico case, we defined as "rural " those respondents from municipalites containing 220,292 residents or fewer, which corresponds to the 1st and 2nd quantiles of municipalities across which the Mexican population is distributed (following our sampling protocol outlined in Appendix Section 2.1). In Indonesia, the geographic data we collected are not finely scaled enough to identify urban and rural respondents. Note that costs shown do not include incentives.

WEIGHTING

For all three samples, we then used iterative proportional fitting, or raking, to create weights for all respondents according to the distribution of the national populations across gender, education, age cohort, and geography.[20] In Kenya we collated the marginal distributions of Kenyans age 18 years and older in the population using the 2019 census data for the following categories: age (18-29, 30-49, 50-59, 60+), gender (male/female), education (primary or less, secondary, technical training, university or above), and geography (urban/rural) (KNBS 2010). In Mexico we used the same age and gender categories as in Kenya, education (none, secondary or less, technical training, university or above), and geography (size of municipality within the four regions of the country) (INEGI 2015). In

[20]A comparison of weighting procedures suggests that choosing appropriate weighting variables is more important than using a more complex statistical procedure, and that raking works just as well as more complex methods to reduce bias (Pew Research Center 2018).

Indonesia we used similar age categories, except that the lowest category ranged from 21-29 because we were only able to recruit respondents above the age of 21 in Indonesia. We also weighted on gender (male/female), education (no secondary degree, secondary school graduates, post-secondary degrees, and geography (province). We created the weights using the rake function in the **survey** package in R (Lumley 2020), which iterates to proportionally fit weights based on the marginal distributions of demographic variables of interest. We weighted respondents to fit the national population's marginal distribution of age, gender, education, and geography. In all three cases, we trimmed weights by setting the minimum weight to the 5th percentile and the maximum weight to the 95th percentile.[21]

[21] In Mexico, 512 individuals (9.9% of the sample) had untrimmed weights falling outside this range (0.1-4.6) and were adjusted. In Kenya 67 observations (5% of the sample) had original weights outside of this range (0.7-2.5) and were adjusted.

References

Berinsky, Adam, Michele Margolis, and Michael Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *AJPS* 58 (3): 739–753. ISSN: 1540-5907. https://doi.org/10.1111/ajps.12081.

Berinsky, Adam, Michele Margolis, Michael Sances, and Christopher Warshaw. 2019. "Using screeners to measure respondent attention on self-administered surveys." *Political Science Research and Methods,* 1–8. ISSN: 2049-8470, 2049-8489. https://doi.org/10.1017/psrm.2019.53.

INEGI. 2015. *Population.* Accessed: 20 August, 2019. http://en.www.inegi.org.mx/temas/estructura/default.html#Tabulados.

KNBS. 2010. *2019 Population and Housing Census.* Accessed: 9 June, 2020. http://www.knbs.or.ke.

Latin American Public Opinion Project. 2017. *Americas Barometer, 2016/17.* https://www.vanderbilt.edu/lapop/ab2016/AmericasBarometer_2016-17_Sample_Design.pdf.

Lumley, Thomas. 2020. *survey: analysis of complex survey samples.* R package version 4.0.

Pew Research Center. 2018. *Comparing Survey Sampling Strategies: Random-Digit Dial vs. Voter Files.* https://www.pewresearch.org/methods/wp-content/uploads/sites/10/2018/10/REPORT-Comparing-RDD-and-RBS-Surveys-Final-05.06.19-2.pdf.

Tversky, Amos, and Daniel Kahneman. 1981. "The framing of decisions and the psychology of choice." *Science* 211 (4481): 453–458.