Survey Sampling in the Global South Using Facebook Advertisements

Leah R. Rosenzweig*†, Parrish Bergquist‡, Katherine Hoffmann Pham§, Francesco Rampazzo¶ and Matto Mildenberger‖**

April 23, 2024

Abstract

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. Post-stratification 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.

Word count: 9,729

*Market Shaping Accelerator, University of Chicago. E-mail: lrosenzweig@uchicago.edu

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

‡Department of Political Science, University of Pennsylvania. E-mail: pberg@upenn.edu

§Department of Technology, Operations, and Statistics, NYU Stern School of Business

¶Department of Sociology, Leverhulme Centre for Demographic Science, and Nuffield College, University of Oxford

‖Department of Political Science, University of California Santa Barbara

**The authors thank Michaël Aklin, André Grow, Peter Howe, Anthony Leiserowitz, Jennifer Marlon, Umberto Mignozzetti, Blair Read, and Baobao Zhang 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, and the Yale Program on Climate Change Communication.

1 Introduction

Survey research in the Global South traditionally requires large budgets and lengthy fieldwork, for which researchers hire local enumerators to conduct face-to-face surveys with respondents. Historically, there have been few alternatives to in-person recruitment due to poor electricity coverage and limits to phone and 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% of the world's population had access to mobile internet (GSMA 2023; Rotondi et al. 2020). This growing connectivity presents an opportunity for researchers with limited budgets and those seeking to collect data in settings where the traditional resource-intensive in-person contact model of research is impossible, such as due to natural disasters, violent conflicts, or pandemics.

This paper evaluates one emerging method to recruit online samples: Facebook advertisements. 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. Given this massive user base, the Facebook platform potentially offers researchers access to nationally, culturally, and demographically diverse populations around the world. 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 and Zhou 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). Yet, a longstanding tradition in survey research cautions against an uncritical adoption of this new recruitment method without considering possible limitations on the quality of the survey samples collected (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 using Facebook advertisements to recruit survey samples in the Global South.

While most research has used Facebook to recruit specific target populations, an emerging literature has begun to assess if the platform can also recruit nationally representative survey samples. As Boas, Christenson, and Glick (2020) demonstrate in India and the United States, Facebook's high penetration and the diversity of its user base make the platform an attractive opportunity for recruiting nationally representative samples. Zhang et al. (2020) demonstrate the utility of the platform for doing so, 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 variation in researchers'

targeting strategies influences the cost and representativeness of samples recruited in the UK, Turkey, Spain and the Czech Republic. This work has shown Facebook to be a promising opportunity for recruiting survey respondents, but it leaves us with a geographically incomplete and conceptually unsatisfying understanding of the benefits and drawbacks of using the platform for survey sampling.

We expand the geographic scope and conceptual underpinnings of scholars' understanding of Facebook as a tool to recruit survey respondents. Specifically, we evaluate the representativeness of Facebook-recruited samples in low and middle-income countries, whereas the extant literature has focused primarily on the U.S. and Europe. Due to lower levels of internet access, literacy, and Facebook marketing investment in the Global South, it is not clear that findings from studies fielded in the U.S. and Europe will generalize. We evaluate this question empirically by comparing the demographic composition and estimates of political attitudes from Facebook-recruited samples in Kenya, Mexico, and Indonesia with sample composition and public opinion estimates derived from nationally representative benchmark data sets in each country. We also make these comparisons between our Facebook sample in Indonesia and an online sample recruited by a professional survey firm, which is the most viable 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 survey samples recruited on Facebook, exposes which sources of bias researchers can control, and illuminates how researchers can reduce bias.

Using quota-based sampling, implemented through the Facebook advertising platform, we recruit respondents in Mexico (n=5,168), Kenya (n=1,530), and Indonesia (n=2,829). We show that quota sampling can help to overcome the bias introduced by using the Facebook user base as a sampling frame for national populations. This is especially true in comparison with (counterfactually) relying on the ad platform's optimization algorithm. We also show the noisiness of the demographic data underlying the ad platform and highlight the importance of applying post-stratification weights based on self-reported demographic data. We show that these weights can help to ameliorate representational biases such as, in our case, the over-representation of highly educated individuals. As an initial test of the face validity of our samples, we use a canonical survey experiment to show that Facebook-recruited respondents exhibit a universal behavioral bias. We find that Facebook-recruited individuals report being more politically engaged and more supportive of environmental protection than those recruited by in-person surveys. We suggest that these differences may be due in part to survey mode and in part to sample composition. In terms of practical considerations, we find that the Facebook platform allows for quickly and cheaply recruiting respondents. 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 to

access diverse communities across the Global South. We also argue that 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.

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

We begin by developing a definition of survey quality which builds upon a growing body of research assessing 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 online samples used to draw conclusions about the U.S. public. Scholars typically assess the external validity of U.S. samples by comparing demographic statistics to the U.S. Census (Berinsky, Huber, and Lenz 2012; Huff and Tingley 2015; Coppock and McClellan 2019; Zhang et al. 2020). Assessing survey quality of online samples is relatively straightforward in the US, where high-quality census data are frequently updated and many probability and quota samples are available as benchmarks, but may be more challenging in other contexts.

Extending this work, we conceptually disaggregate and empirically assess the types of errors that threaten the validity of conclusions drawn from Facebook surveys in three Global South countries. We define error similarly to its definition in a regression framework: error represents unobserved disturbance that influences a statistical quantity of interest derived from a survey. Such error is problematic to the extent that it causes estimates derived from our survey to differ systematically from true parameters in the population of interest (i.e., estimands). We thus focus on examining bias, defined as the degree of systematic difference between estimates derived from our survey and the true parameter in the population. By disaggregating the sources of error that threaten validity, we provide guidance for scholars to consider in deciding whether to use Facebook as a survey recruitment tool.

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-based surveys. Since the introduction of this framework, survey

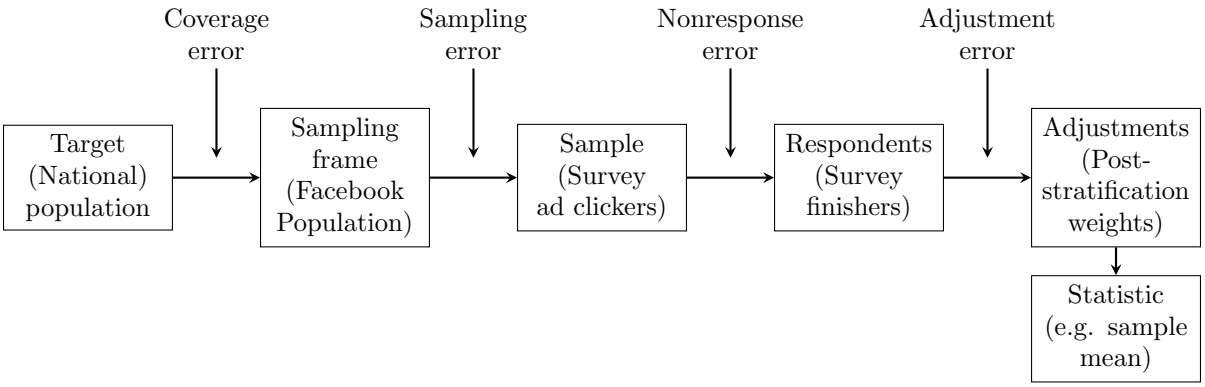

Figure 1: **Components of Representation Error.** The figure disaggregates the threats to external validity in survey samples and defines the survey error framework in the context of our study: using Facebook to recruit a sample that is representative of a national population.

methodologists have defined two groups of quality indicators that can be applied to statistics derived from surveys: measurement error[1] and representation error (Groves et al. 2011; Groves and Lyberg 2010). Here we focus on quality concepts that pertain to representation. Errors of representation are systematic or random imperfections in the relationship between a target population, sampling frame, and sampling units. Systematic errors of representation contribute to threats to external validity.

Figure 1, which is adapted from Groves et al. (2011) and Groves and Lyberg (2010), shows the inferential steps required to draw conclusions about a population from a set of responses in a sample-based survey. The figure also includes the instantiation of each concept in our case: survey samples recruited from Facebook. We will focus on the errors that arise at each of these inferential steps.[2] The first error that we are concerned with is coverage error, defined as the gap between a target population and a sampling frame. In our case, the target population is the national adult population in each country we study, and the sampling frame is the national adult population that has a Facebook account in each country. To the extent that there are adult residents of each country who are not on Facebook, we will have "undercoverage." If our quantity of interest is a descriptive statistic, such as the mean value on a political attitude question, coverage bias is quantified as the difference between the mean value in the national population and the mean value for the Facebook population. Of course, we cannot directly observe average political attitudes for the national population; instead we infer this quantity from nationally representative, in-person benchmark surveys. The second

1. Measurement addresses gaps between the constructs that a survey is designed to assess, the measures used in a survey, and responses to those measures. These types of errors present threats to internal validity. Measurement error arises from systematic or random imperfections in the relationships between constructs, survey questions, individuals' responses to those questions, and researcher processing of responses. While a crucial research consideration, measurement error is primarily a matter of questionnaire design. In general it is not platform-specific and, thus, is not the focus of this study. Still, while we do not provide a full treatment of measurement error, we do examine one component of measurement error–respondent attention–in SI Section S2.3.1. This is to satisfy the curiosity of readers who might wonder whether this component of measurement error might affect the quality of Facebook samples in ways that are discernibly different from samples drawn from other platforms.

2. Although this framework is presented linearly here, these sources of error do not necessarily sum to a whole.

source of representation error is sampling error, which arises because not all individuals in the sampling frame are surveyed. Random sampling variance arises because many different sets of individuals could be drawn from the sampling frame, simply by chance. Sampling bias would arise if different individuals in the sampling frame have different chances of being included in the sample. The third source of representation error is non-response error. Here we focus on unit non-response error, which arises when some sampled individuals fail to record high-quality, complete responses to a survey.In this context, 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. The final source of representation error is adjustment error, which 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 them (Bailey 2024).

We examine total survey error and disaggregate its components with two sets of empirical analyses. Our analyses build on prior work that assesses bias by comparing survey-derived statistics to external benchmarks, such as a "gold-standard" survey, census data, 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). 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 the types of error that are under researchers' control.

3 Data Collection

3.1 Case selection

To ensure that our analysis is broadly useful to researchers, we fielded surveys in three distinct regions of the Global South. We selected Mexico, Kenya, and Indonesia – 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 with high literacy rates and mobile phone access to ensure the 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

3. We calculated these figures based on the number of users per country that Facebook reports (Araujo et al. 2017) and the population figures for individuals age 15 years and older in each country available from the UN (United Nations 2019; Ševčíková

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), meaning that a majority of citizens would be able to read and self administer an online survey (World Bank 2020).

3.2 Benchmark data sets

We assess the quality of our Facebook-recruited samples by comparing statistical summaries derived from them with benchmarks designed to be representative of the national population in each of our countries of interest. In all three countries, we use the national census as one benchmark. We also use well-respected, in-person, nationally representative surveys fielded in each country. These surveys are 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 commercial survey firm that recruits respondents through its online panel. We make this comparison in order to provide insights about the comparative quality of Facebook samples with the most viable alternative for most researchers considering online surveys. Dates of data collection for each data set are presented in Table 1.

3.3 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 as closely as possible those used by these benchmarks. For Indonesia, we used a stratified sample by province. Next, 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).

2020).

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

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

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 (e.g. states). In Mexico and Indonesia, we targeted respondents according to these predefined geographic entities: municipalities in Mexico and provinces in Indonesia. We are therefore able to examine the viability of both approaches to geographic targeting. Table 1 summarizes the sampling strategy for each country. Section 3.3 of the Supplementary Information (SI) includes further details about the approach taken in each country, along with an extended discussion of the constraints and opportunities embedded within Facebook's advertising platform.

Table 1: Comparison of the 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 ($\sim$ \$0.50)	Yes ($\sim$ \$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	1,530	3,277

** The parentheses denote strata added partway through data collection to correct for observed sample imbalances.*

3.4 Survey instruments

To direct respondents to the surveys, we created Facebook pages representing our survey campaigns, and placed ads from these pages to target people within the sampling strata described. After clicking on the Facebook ad, respondents were sent to a survey hosted on Qualtrics. Respondents provided informed consent before agreeing to participate in the survey. In Mexico the survey was administered in Spanish. In Kenya,

6. We used the Afrobarometer sampling clusters geolocation. One reason why we did not target ads based on administrative unit is because Facebook only had provinces, the outdated administrative unit in Kenya, not the more recent 47 counties introduced in 2010.

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. In Indonesia, the survey was offered in Bahasa. Upon completing the survey, respondents were directed to a thank-you page with an embedded Facebook "Pixel" which allowed Facebook to identify the users who clicked on the ad and completed the survey.[7]

We designed all surveys to collect information on demographics and attitudes that we could compare to each country's census and to our benchmark surveys. The survey used in Kenya 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 survey used in Mexico replicated certain questions from the Mexican Census and the LAPOP survey, fielded in early 2019 (LAPOP 2018–2019). In Indonesia we replicated survey questions from the Asian Barometer and a survey implemented by Dynata in October 2021. Full copies of our surveys are included in our replication file in the Harvard Dataverse.

3.5 Poststratification adjustments

For all samples, we then 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 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 SI Section S6.

4 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 those reflected in the benchmark surveys and the national census in each country. Our quantity of interest is the weighted mean survey response derived from our sample.

Figure 2 plots the distribution of several demographic characteristics, compared with the benchmark surveys and national census in each country. In all three countries, the weighted Facebook samples (filled, dark blue squares in Figure 2) differ most from the national population (gold crosses in Figure 2) with respect to age and education. In Kenya and Indonesia, the Facebook survey samples are younger than the national

7. The Facebook pixel tracking survey completions did not successfully register completes in the Indonesia survey. For this reason we anticipate that our costs per results represent a higher estimate than may have been achieved had Facebook had the information about which types of people were more likely to complete the survey after clicking on the ad, since this would allow the algorithm to optimize ad targeting with that information. This could lead to more efficient ad spending/high conversions, but might also imply a slightly different sample composition. We did use the pixel in the case of Kenya and Mexico and see similar results in terms of over-represented subgroups in all samples.

population and the barometer samples. (We note that the Asian Barometer actually under-represents young respondents.) By contrast, the Facebook sample in Mexico contains a greater share of respondents 50 years and older than the national population, and the LAPOP sample shows a similar bias.

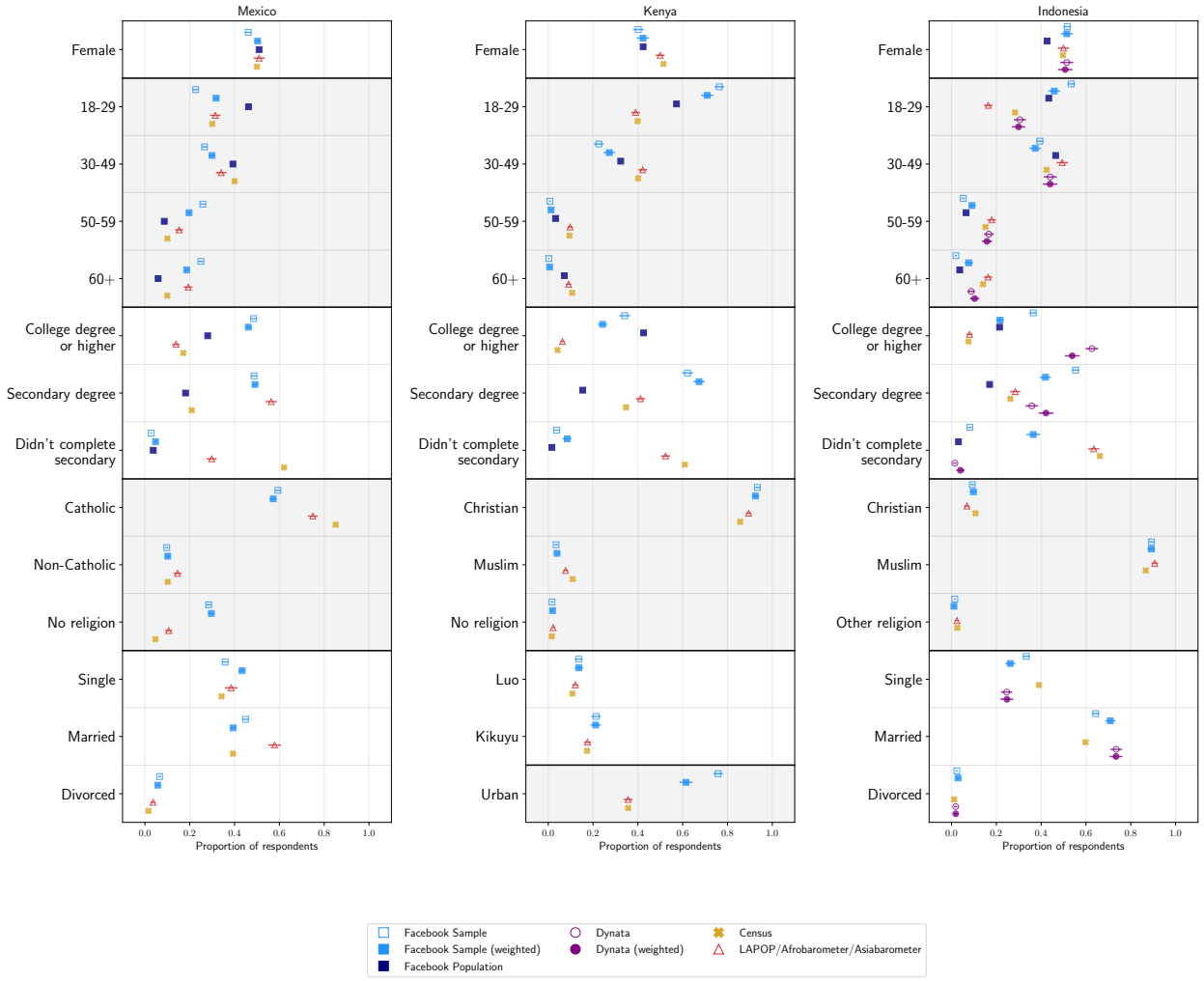

Figure 2: Comparison between demographics reported in the national census, nationally representative survey samples (LAPOP/Afrobarometer/Asian Barometer), the Facebook population, and our Facebook sample for Mexico, Kenya, and Indonesia. The Facebook sample is weighted using raking to match the national census on gender, education, age cohort, and geography. Mexican comparison data are from INEGI (INEGI 2015) and LAPOP. Kenyan comparison data are from the national census (KNBS 2010) and the Afrobarometer (Afrobarometer 2016). Indonesian comparison data come from the Indonesian Census, the Asian Barometer, and the survey recruited with Dynata.

In general, the Facebook samples are more educated than national populations and the barometer surveys. However, the Facebook sample in Indonesia is actually less biased with respect to 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. The commercial online sample also under-represents

lower education respondents even more severely than our Facebook sample. In this important respect, the Facebook sampling method in Indonesia performs quite well relative to the Dynata sample (purple circles in Figure 2).

There is a relatively little bias observed in terms of gender in 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 context - compared to 36% reported in the census.

4.1 Coverage error

To examine the specific sources of these biases, we begin with coverage bias which is 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 represent the demographic characteristics of the Facebook population.

Coverage bias seems to contribute to the over-representation of more highly educated respondents that we observe 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, coverage bias with respect to education is even larger than overall bias (the difference between the census and our weighted Facebook estimates), and it appears that quota sampling has alleviated this bias. In Mexico, compared with the national population, college-educated individuals are even more over-represented in the Facebook sample than in the Facebook population. Coverage bias may account for some of the over-representation of more educated respondents in our samples, but sampling and non-response error are also contributing factors, as we discuss in the next sections.

Coverage bias also helps to 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. However, coverage bias does not fully account for the bias in the samples; in Kenya and Indonesia the Facebook samples over-represent 18-29 year olds even compared with the Facebook population.

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 most closely matches the census in Mexico, 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 of females present in the Facebook population. In Indonesia, however, the Facebook sample (weighted and unweighted) closely approximates the gender balance in the national population. Here again, this is likely due to more precise closing of ads after quotas were reached.

To assess how much of an issue coverage bias might be for particular demographics of interest, 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 post-stratification weighting can help mitigate coverage bias even in cases where the sampling frame may not be representative of the population on a particular dimension.

4.2 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 to recruit 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). There is always a random, ignorable component of sampling error. Sampling bias can occur if some people in the sampling frame are systematically excluded or underrepresented, for example, if some people have zero chance of being included in the sample. Sampling bias could arise at two points in the sampling process. First, the design of the ad may appeal to some individuals more than others. Researchers have control over this step, and can, for example, experiment with the design of the ad 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 as described in Section 3.3, but whether the ads actually reach the targeted individuals is determined by Facebook's back-end data and algorithms, which are beyond researchers' control. Here we discuss the challenges associated with precisely sampling individuals based on their demographic characteristics, and propose diagnostic steps that researchers can take to minimize

sampling bias.

To assess sampling bias in our study, we focus on the group of individuals who clicked on our survey ad. We call these individuals "ad clickers," to distinguish them from the sample of respondents who completed the survey and whose demographic information is reflected in Figure 2. Using the Facebook ad platform allows us to observe some demographic information for ad clickers, even if they did not complete the survey. Demographic information is assigned to each individual according to the sampling stratum through which they received the ad.

We use Facebook-inferred demographic information to check whether we were able to recruit at least one individual from each stratum and, conversely, whether any type of targeted individual was systematically excluded. In Kenya and Mexico, we were able to do so, although in Kenya we did not reach the target number of individuals for 49 of our 66 strata.[8] In Indonesia, we were unable to recruit any respondents from 67 (25%) of the 272 strata that we targeted. The vast majority of these strata contained men and women over 50 years old, though we also failed to reach any women between 30 and 49 years old in three provinces.

Of course, if Facebook's back-end data are inaccurate, then even perfect success at recruiting people from the strata we designed does not ensure that every individual in the population has a chance of being included in the sample. 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.

The accuracy of Facebook's advertising targeting varies across demographics and between countries.

8. The strata targets were set according to the Afrobarometer's population-derived weights associated with each stratum's geolocation. The number of respondents targeted per stratum ranged from 4 - 167. The strata that fell short were missing a median of three respondents (Min: 1, max: 52). 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.

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%	64%	55%
Additional criteria:			
Education level	30%	12%‡	

** This reflects the proportion of respondents who were at least 32 years old, given that they responded to an ad targeting this age group specifically. 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.*

Facebook's targeting was remarkably accurate in correctly identifying respondents' gender in Mexico and Kenya (99% and 91% match, respectively), but slightly less so in Indonesia (76%). In both Indonesia and Kenya, the 10-20% of respondents who were recruited from an ad that was targeted 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 the incentive. 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), whereas in Kenya fewer than half of our respondents (44%) reported an age that was consistent with the age bracket that Facebook assigned to them. Geographic targeting was less accurate in all three countries. Furthermore, where we used it (in Kenya and Mexico), targeting by education was very imprecise. In SI Section S1, we report more details on these comparisons for interested readers.

These findings indicate that inferred demographics are not always accurate. Therefore, researchers should also apply weights based on self-reported demographic data. We also recommend that researchers periodically examine the composition of the sample of ad clickers, in order 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 our ads in each country. Without specified quotas, the Facebook ad platform would simply show the ad to the cheapest (ie, most common or easily engaged) types of respondents. 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. Here the goal would be to increase entrants from under-represented groups in the sample of ad clickers. To encourage more people from these groups to enter the survey, researchers could increase spending for the ad sets targeting these groups, modify ads to make them more attractive to these groups, or leave these ads running for a longer period.

Despite these limitations, 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 and thus, usually,[9] who are most prevalent in the Facebook population. Our disaggregation of the ad campaign across quota sampling cells overrides this regression towards the most common (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 we cannot fully overcome the gap between the Facebook population and the national population with quota sampling.

9. Intuitively, the cheapest respondents should be the ones who are most prevalent on the platform, but recall that researchers are competing with other ad bidders. If people in certain strata are subject to a lot of advertising competition, they may be more expensive to recruit (even if they are common on the platform). Similarly, if people in certain strata are more likely to engage with ads (by clicking on them and/or converting as measured by Facebook's pixel), they may be cheaper to recruit (even if they are less common on the platform).

Nevertheless, we can aim to further mitigate remaining biases with weighting.

4.3 Non-response error

Unit non-response is quite common across survey contexts, and it causes bias if it is non-random and correlated with the attributes or opinions measured in a survey (Bailey 2024). We examine whether non-response is systematic using the same Facebook-assigned demographic information that we used for the analysis shown in Figure S1. We assess whether the likelihood of entering but dropping out of the survey was correlated with users' demographic characteristics (assigned by Facebook). Figure 3 shows the results from regressions of attrition on Facebook-assigned gender and age groups for our samples. Here we define attrition and non-response interchangeably, as entering the survey but failing to provide high-quality and complete responses. In Mexico, this includes those who entered but failed to complete the (very short) survey. In Indonesia and Kenya, this includes those who entered but failed to complete the survey and those who completed the survey in under five minutes are defined. The regressions show that attrition is systematic, but the predictors of attrition vary across samples.

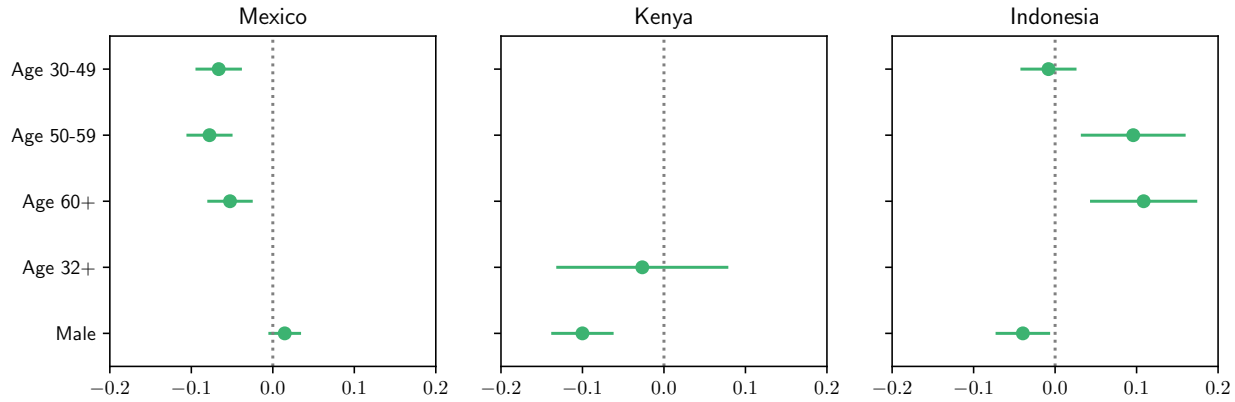

Figure 3: **Predictors of non-response:** The figure shows the results from a linear regression of attrition (attrition=1, completion=0) on the demographic characteristics used by Facebook to target individuals and invite them into our survey. The omitted categories are Female and 18-29 for the regressions using the data from Mexico, Female and 21-29 for the regressions using the data from Indonesia, and Female and younger than 32 years old for the regression using the data from Kenya. 95% confidence intervals are reported based on heteroskedasticity-robust standard errors.

Incorporating age and gender into the weights we use to adjust our samples helps to alleviate non-response bias associated with these observable characteristics. This is because a higher likelihood of attrition leads to under-representation in the sample (and vice versa). For the observable characteristics we use for weighting, we reduce non-response bias by upwardly weighting individuals with under-represented characteristics and downwardly weighting those with over-represented characteristics.[10] Additionally, researchers could adjust

10. Note that this correction is imperfect, since our analysis of non-response bias is based on Facebook-assigned demographics,

their sampling strategy on the fly to improve response rates among high-attrition groups. Researchers could periodically examine attrition across the demographic groups used for targeting, and adjust the survey to improve response rates among groups with high rates of attrition. 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), non-response weights (Sun et al. 2018), or other methods Bailey (2024).

4.4 Adjustment Error

While we can minimize bias due to non-response by weighting on observable respondent characteristics that predict non-response, it is possible that the use of post-stratification weights exacerbates some of the other biases we have diagnosed, rather than ameliorating them. This would reflect adjustment bias. We examine adjustment error by first examining whether the weights correct demographic imbalances in the sample, and we find that our (trimmed) weights ameliorate biases. We then investigate whether the weights introduce bias on other variables we measure, and we find that they do not.

Examining Figure 2, we can observe whether the application of weights corrects demographic imbalances between our Facebook samples and the national populations by comparing the unweighted (hollow, light blue squares) and weighted (filled, light blue squares) Facebook samples with the census data (gold crosses). Weighting corrects the slight imbalances with respect to gender, 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 the trimming of our weights at the 95th percentile. We do this to avoid excessively over-weighting very rare respondent types. The Facebook sample almost perfectly matches the census populations if we do not restrict the weights.

While weighting improves the representativeness of the sample on the demographic characteristics that we incorporate into the weights, the application of weights could introduce bias if it draws the sample distribution away from the population distribution on other dimensions. To assess whether this is happening, we investigate the extent to which weighting the sample reduces the accuracy of descriptive inferences about demographic characteristics that are not incorporated into the weights (shown in the bottom sections of Figure 2). Weighting minimally affects the distribution of the demographic variables that are not incorporated into our weights (religion, marital status, tribe) with the exception of the urban bias in the Kenya sample,

whereas our weights are based on self-reported demographics.

which is reduced. This suggests a negligible correlation between age, gender, education, and geography and the other demographic variables and, correspondingly, provides some assurance that weighting does not introduce bias into descriptive inferences. Likewise, weighting only slightly impacts our estimates of public opinion (analyzed below, in Section 6) and tends to move our public opinion estimates slightly towards the benchmark estimates, rather than away from them. Overall, our post-stratification adjustments improve the quality of the statistics derived from our surveys, without introducing additional bias.

5 Replicating classic survey-experimental findings

For most social science researchers, the quantities of interest from a survey are measures of 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. Supplementary Section S4 describes the experiment in detail. 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) among a US-based MTurk sample, and results from our Facebook samples in Kenya and Indonesia.[11] Tversky and Kahneman's results replicate in both samples: respondents exhibit loss aversion, as predicted in the classic experiment.

Table 3: Replication of Tversky and Kahneman (1981) Disease Problem

	Tversky & Kahneman		**Berinsky, Huber, & Lenz**		**Kenya Facebook sample (unweighted)**		**Indonesia Facebook sample (unweighted)**	
Options	Save	Die	Save	Die	Save	Die	Save	Die
Certain	72%	22%	74%	38%	62%	36%	53%	46%
Risky	28%	78%	36%	62%	38%	64%	47%	54%

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.

6 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 which is of increasing interest to social scientists. In each case we compare responses from our Facebook samples to the same opinion questions fielded in benchmark surveys (the Afrobarometer, the LAPOP, and

11. The Indonesia data analyzed here is from the pilot version of the survey, fielded in July 2023. SI Section S4 provides more information on the pilot wave of the survey.

the Asian Barometer).[12] Figure 4 plots the Facebook sample estimates, both weighted and unweighted, as well as the benchmark estimates. In general, where there are gaps between the estimates derived from our Facebook samples and those reported by the barometer surveys, respondents from our Facebook samples report greater political activity than respondents from the benchmark surveys. Note that the application of weights tends to move the Facebook-derived estimates towards the estimates derived from the benchmark surveys. This reinforces the point made in Section 4.4, that we see little evidence of adjustment bias. Instead the accuracy of our estimates tends to improve with the application of weights.

The gaps we observe are likely attributable to some combination of sample composition and survey mode. Representation bias could account for some upward bias in our estimates of political activity, since our Facebook samples over-represent highly educated respondents who are likely to be more politically engaged (Verba, Schlozman, and Brady 1995). Still, we suspect that survey mode also plays a role here and that the barometer survey data reflect under-reporting of certain political activities. Social desirability bias may deter respondents from reporting engagement in 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) due to social desirability bias. In both countries, the only place where Facebook samples report *lower* activity than the corresponding barometer estimate is with respect to voter turnout. Moreover, the Facebook sample estimates are closer to the official turnout figures reported in each country (indicated by the green crosses in Figure 4). Likewise, 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.[13] Overall, to the degree that survey mode affects responses, it is not clear that the barometers necessarily enable more accurate data collection on measures of political activity, although we lack measures of ground truth with which to verify this (aside from voter turnout).

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

12. The 8th round of the Afrobarometer survey was conducted in person in Kenya just a few days prior to when we fielded our Facebook survey. At the end of our survey we asked if respondents had been previously surveyed by any other organization. While 47% reported that they were surveyed by census enumerators, less than 1% reported being surveyed by the Afrobarometer. 16% said they were surveyed but did not recall by whom.

13. To be sure, since Dynata is also an online survey, the higher estimate of voter turnout derived from the Dynata sample (compared with Facebook) complicates this explanation. The higher level of education in the Dynata sample may counteract the mode effect.

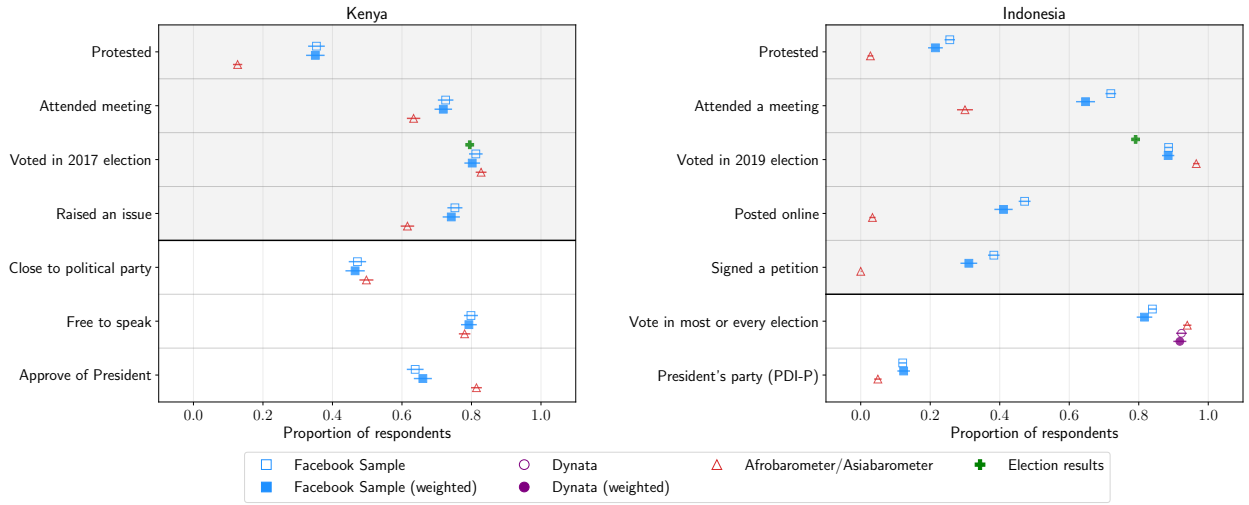

Figure 4: This figure shows responses from Afrobarometer/Asian Barometer and Facebook samples in Kenya and Indonesia, to questions related to political attitudes and behaviors. It compares sample respondents according to self-reported behaviors including identifying with a political party, voting, and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

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 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, and Asian Barometer respondents were more likely to report that they did not know the answer to these questions. (8-17% of the Asian Barometer respondents responded that they did not know whether they had participated in these activities, compared with 1-2% of Facebook sample respondents.) Given how memorable these activities can be, the high rate of "don't know" responses reported in the Asian Barometer survey suggests that individuals may be reticent to admit to having done them. 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 accurately assessing policy views among the public, we compare opinions about environmental policy and climate change in Mexico and Indonesia. To measure policy views in Mexico, we replicated the question wording and response options from a question that LAPOP asked in its 2019 survey of Mexican respondents (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 the respondents recruited from Dynata's online panel. This provides a comparison between a Facebook-recruited sample and the most viable alternative for most researchers. Facebook respondents are 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).

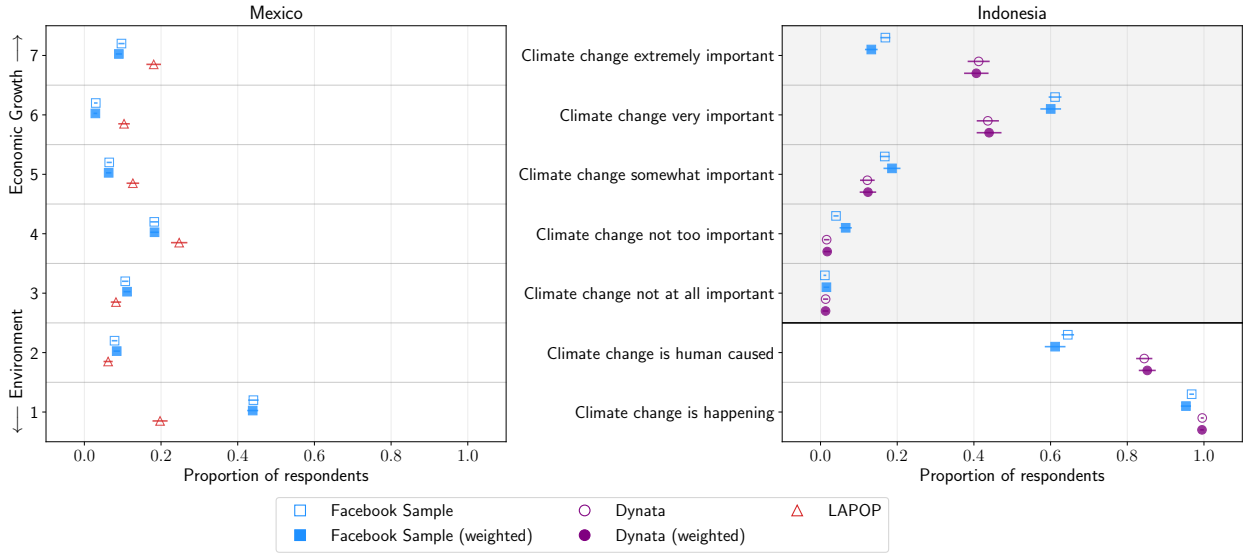

Figure 5: This figure shows responses from LAPOP and Facebook samples, to the question of whether environmental protection (1) or economic growth (7) should be given priority. Results are weighted for the Facebook sample, according to the procedure described in section 3.5, and unweighted for the LAPOP sample, consistent with the survey design documentation.

7 Survey costs

Researchers typically weigh survey quality concerns against the relative costs of different survey tools. For Facebook-recruited surveys, these typically include the advertising costs for using the platform and optional incentive costs associated with paying respondents. We explore the extent to which Facebook-recruited surveys are cost effective compared to other modes of data collection. Supplementary Section S5 provides details about the reach of our survey as advertised on Facebook, and the cost of reaching different demographic

groups.[14]

Sampling on Facebook is quite cost-effective. Not including incentives, the mean cost per completed survey was $0.16 in Mexico[15], $0.85 in Kenya,[16] and $0.92 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, these 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 Dynata sample (n=1,130) in Indonesia cost $5.75 per completed survey. In-person surveys are even more expensive to field.

8 Conclusion

The Internet has enabled more of the world's citizens to find a political voice than ever before, but high-profile prediction failures have also raised new questions about how to accurately measure and represent the global diversity of political views. Facebook, due to its broad user base, offers an opportunity for researchers to quickly and reliably recruit subjects from countries that are underrepresented on existing online subject recruitment platforms. While survey researchers have begun to use Facebook to recruit survey samples, the literature lacks a careful and comprehensive assessment of the quality of Facebook samples recruited in the Global South. In this paper, we have used a series of comparative analyses to scrutinize the quality of Facebook-recruited samples in three countries, highlighting some of the advantages and shortcomings of this method and providing practical guidance to researchers on how to measure – and partially address – these 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 associated with estimates derived from Facebook-recruited surveys. We showed that coverage error favors respondents that have achieved higher levels of education, on average, than the general population in the countries we survey. This over-representation of highly educated respondents is a structural feature of the Facebook platform in our case countries – at least for now. By

14. These costs vary because Facebook ads are deployed using a bidding system in which hard-to-reach populations are more expensive for advertisers to target.

15. These statistics reflect the full campaign, including our low-education oversample. In Section S5 of the Supplementary Information, we include separate data for the initial sample (without education targeting) and for the oversample in which we targeted respondents with a high school degree or lower levels of education.

16. This amount is derived from 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 surveys completed on Qualtrics which were determined to be valid and were used in the analysis.

extension, the demographics of Facebook users may similarly differ from those of the national population in other countries, and researchers should investigate these imbalances before deciding to use Facebook as a survey recruitment tool in a particular context.

We showed that researchers can mitigate bias associated with coverage error by targeting recruitment resources towards under-represented respondents. Quota sampling through the ad platform enabled us to recruit a sample that is more representative than it would be if the advertising algorithm were left to its own devices, and we illustrate how researchers can continuously monitor incoming samples in order to redirect resources to recruit under-represented respondents. However, we found that the back-end demographic data that Facebook uses for targeting are noisy, particularly with respect to individuals' education and location. This implies that quota sampling cannot substitute for the use of design weights based on self-reported demographics. We next demonstrated how researchers can assess non-response rates across quota cells during data collection to inform an iterative sampling strategy that maximizes response rates among high-attrition groups.

Finally, we examined the effect of design weights on the representativeness of our sample – finding that the weights reduce, but do not completely eliminate, demographic imbalances. In part this is due to our decision to trim weights at the 95th percentile, which prevents over-weighting of very rare response types. The weights affect descriptive inferences only slightly, and our weighted estimates tend to be closer to the benchmark estimates, compared with the unweighted estimates. We conclude that adjustment bias is minimal.

We also showed that the costs of using Facebook to recruit survey respondents can be quite low, though costs depend on the targeting strategy used. Naturally, the use of incentives for survey completion influences the cost-quality tradeoff. In Mexico, we successfully recruited respondents without incentives, likely because the survey was quite short and internet penetration is quite high. 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 the survey link. Future research could investigate sampling strategies to leverage the social nature of the platform, for example using snowball sampling that promotes forwarding of the survey link.

Scholars should consider context before deciding to use Facebook to recruit subjects. Using Facebook as a method of respondent recruitment will be particularly successful in contexts where 1) phone and internet penetration is widespread, 2) literacy rates are high, and 3) recent census data are available. The ability to weight Facebook responses is predicated on access to reliable census (or other benchmark population) data. Researchers looking to adopt this method should also keep in mind that the nature of 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 who answer surveys on the platform might have different assumptions about the government's surveillance of their responses. Researchers should consider these concerns in the survey design process, and future research could consider how internal validity varies according to political context.

We do not suggest that Facebook should 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. We have demonstrated a series of iterative steps that survey researchers can use to diagnose coverage error, sampling error, and non-response error in Facebook surveys. Using this information, researchers can iteratively adjust their sampling protocol, model non-response bias, and utilize post-stratification weighting to mitigate sample biases. Overall, Facebook represents a valuable tool that, when used well, promises to open new frontiers in public opinion research.

References

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

Ansolabehere, S, and Brian F Schaffner. 2010. "Does survey mode still matter." *Findings from a.*

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. doi:10.1145/3091478.3091513.

Bailey, Michael. 2024. *Polling at a crossroads: Rethinking modern survey research.* Cambridge University Press.

Berinsky, Adam J, Gregory A Huber, and Gabriel S Lenz. 2012. "Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk." *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." *American Political Science Review* 109 (1): 62–78.

Broockman, David E, and Donald P 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 A McClellan. 2019. "Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents." *Research & Politics* 6 (1): 2053168018822174.

Deming, W 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? Experimental evidence from a new democracy." *American Journal of Political Science.*

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." *American Political Science Review:* 1–16.

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

Groves, Robert M, 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: Advantages, challenges, and practical considerations." *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 S Lenz, Maksim Pinkovskiy, and James M Snyder Jr. 2015. "Voter learning in state primary elections." *American Journal of Political Science* 59 (1): 91–108.

Holbrook, Allyson L, and Jon A Krosnick. 2010. "Social desirability bias in voter turnout reports: Tests using the item count technique." *Public Opinion Quarterly* 74 (1): 37–67.

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.

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.

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

Kilavuz, M Tahir, Sharan Grewal, and Robert Kubinec. 2023. "Ghosts of the Black Decade: How legacies of violence shaped Algeria's Hirak protests." *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 C Matz, Samuel D 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 M Markowitz, Peter D 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." *The American Economic Review* 80 (2): 319–323.

McGovern, Mark E, David Canning, and Till Bärnighausen. 2018. "Accounting for non-response bias using participation incentives and survey design: An application using gift vouchers." *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.

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 M Tahir Kilavuz. 2023. "Regime Support and Gender Quotas in Autocracies." *American Political Science Review:* 1–18.

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

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. doi:10.1177/0894439316666262.

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

Rosenzweig, Leah R, and Yang-Yang 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.*

Ryan, Timothy J. 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: Evidence from survey experiments." *American Journal of Political Science* 58 (1): 212–225.

Santoso, Lie Philip, Robert Stein, and Randy Stevenson. 2016. "Survey experiments with Google consumer surveys: Promise and pitfalls for academic research in social science." *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 Lehman Schlozman, and Henry E 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 D 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

S1 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.

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] 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.

1. 1st quartile: 33, 3rd quartile: 46.
2. For the province-level oversampling of older and less-educated respondents, we set a quota of 5 respondents per stratum.
3. Min: 1, max: 52.

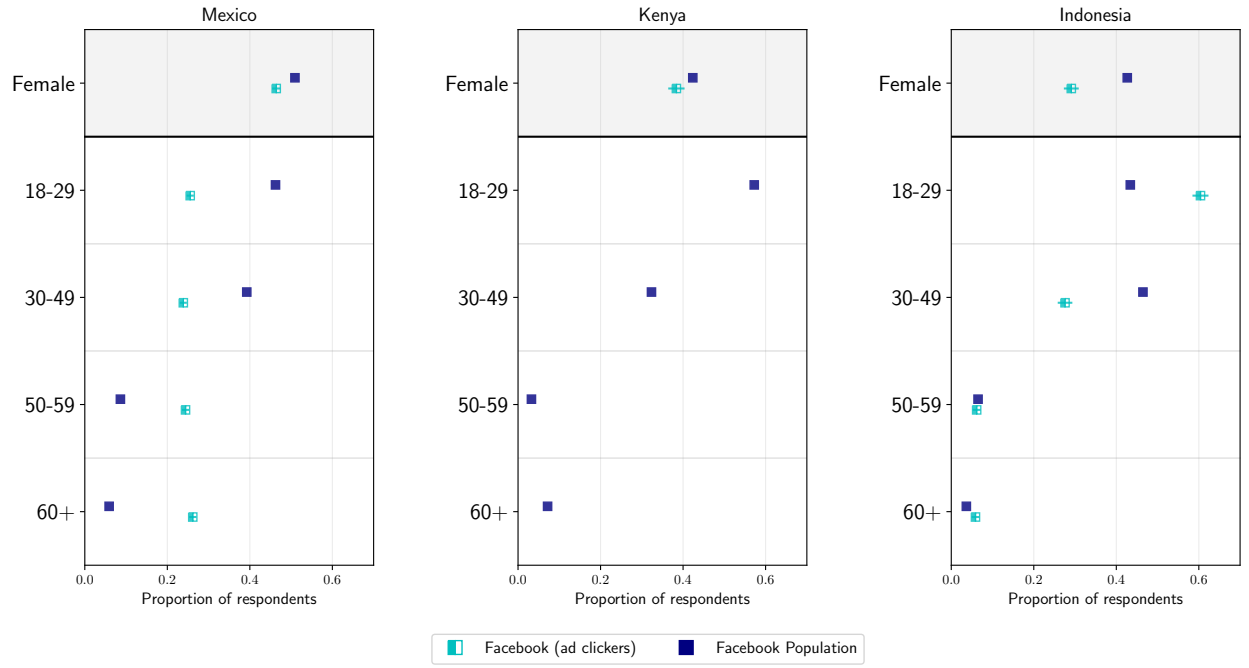

Figure S1: **Sampling error:** The figure shows the proportion of the Facebook population (dark, solid squares) and ad clickers (light, partially filled squares) from each age and gender group used in our sampling quotas. Demographics are based on Facebook's back-end data about users. Note that in Indonesia 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.

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 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 imprecise in both countries where we used it: Kenya and Mexico. In our second round of data collection in Mexico we targeted respondents with a high school degree and lower levels of education. Only 30% of the respondents who reached our survey through these ads self-reported education levels that matched Facebook's categorization. However, 43% of the individuals in the low-education targeted group reported their highest level of education as "Bachillerato/ Profesional Técnico/ Media Superior." Facebook may consider these technical school degrees — which overlap with the type of schooling that U.S. residents might call "high school" — equivalent to high school. If this is the case, then Facebook correctly categorized 73% of the respondents in this group. The ambiguity in Facebook's category definitions for this group illustrates one challenge of using a cross-national platform with universal targeting categories.

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.

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 was slightly less accurate in all three countries. In Mexico, 67% of respondents reported living in a municipality 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 64% match rate between the Facebook target province and the Afrobarometer province. The remainder of respondents in hard-to-reach age or education groups (n=67) were targeted at the province level. In these strata, 69% of respondent-reported towns of residence fell within the targeted province.

S2 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

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.

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.

the predominant unit of administrative organization since devolution in 2010. The availability of county targeting does not seem to be an issue 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.

S2.1 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

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.

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 × 4 municipality size 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.

S2.2 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 × 2 genders × 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 (∼ $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.

S2.3 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

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 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.

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

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

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 ($\sim$ 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 ads targeting the province level, we excluded each province's

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

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 the survey, did not report their age or gender, or finished the survey in less than five minutes, as well as removing duplicate entries, out sample contained 1,530 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.

S2.3.1 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.

of data sources including a user's location, websites visited, pages liked, and information posted on the user's profile page.

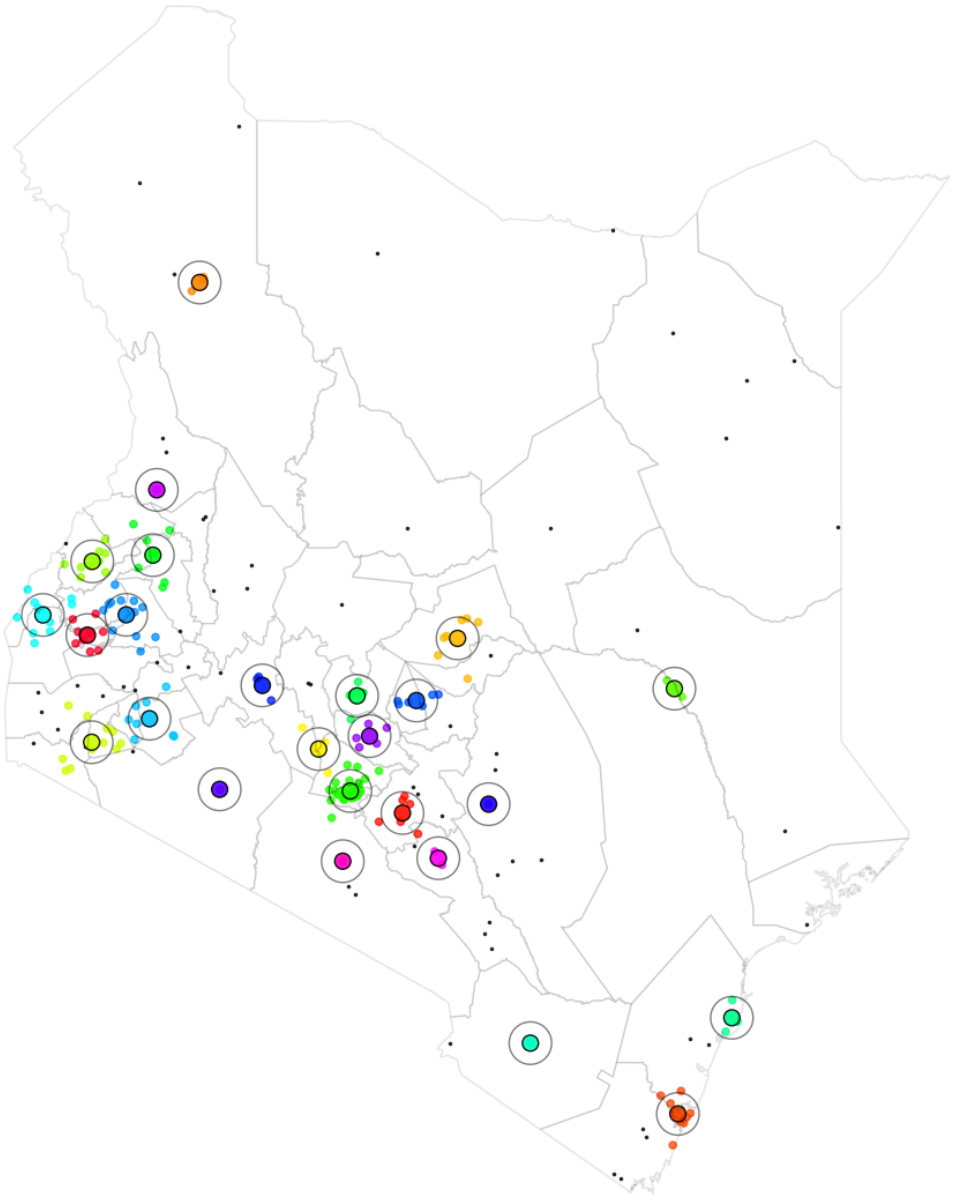

Figure S4: This map shows how Afrobarometer clusters were grouped together using the K-means algorithm. The colored dots with black outlines represent the 25 final K-means centroids used to target Facebook ads in the full study. The black circles represent 20km radii about these centroids. The colored 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.

S3 Poststratification adjustments

As mentioned in the main text, we used iterative proportional fitting, or raking, to create weights for respondents according to the distribution of the national populations across gender, education, age cohort, and geography.[20] 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.

S4 Prospect Theory Experiment

The surveys in Kenya and Indonesia included a replication of Tversky and Kahneman's (1981) "disease problem."

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. Assume that the exact scientific estimates of the consequences of the programs are as follows.

Then respondents are presented with two policy options: a single certain option, and a risky option that assigns probabilities to two different outcomes. In expectation, the payoffs of both policies are equal, so differences in respondents' choices should be driven primarily by their appetites for risk. Respondents are randomly assigned to one of two conditions: either the policy options are framed in terms of a loss (number of deaths) or the policy options are framed as a gain (number of lives saved). The canonical finding that has been replicated in many samples across the globe suggests that people are loss averse — favoring the certain option when it is framed in terms of lives saved, and the risky option when it is framed in terms of lives lost (Tversky and Kahneman 1981).

S5 Survey Costs

Table S2 shows the reach and total spending of our survey campaigns.

Table S2: Reach of survey

	Mexico	**Kenya**	**Indonesia**
Impressions	492,564	649,264	7,153,465
Reach	439,056	318,960	4,157,815
Clicks	24,314	8,206	23,815
Survey results	5,168	1,530	3,277
Total spent	$847.76	$1,293.91	$3,000
Click through rate (%)	0.049	0.013	0.006
Completion rate (%)	0.010	0.002	0.0005
Cost per click ($)	$0.03	$0.16	$0.13
Cost per completed survey ($)	$0.16	$0.85	$0.92

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. These survey results are those reported by Qualtrics. The click through rate and the completion rate are defined with respect to impressions.

A key question for researchers is whether Facebook surveys are cost-effective along different dimensions of cost-effectiveness. When implementing Facebook surveys, there are two types of marginal costs: (1) advertising costs and (2) costs associated with paying respondents (as we did in Kenya). Here, we focus on advertising costs. We quantify advertising costs in terms of cost per click and cost per completed survey.

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 (Mercer, Lau, and Kennedy 2018).

Table S3: Cost of Facebook targeting

	Mexico			Kenya		
	Median	Max	Min	Median	Max	Min
Click through rate (%)	5.56	13.25	2.16	1.2	6.0	0.7
Completion rate (%)	1.24	4.29	0.29	0.2	2.8	0.0
Cost per click ($)	0.03	0.066	0.02	0.15	0.47	0.02
Cost per completed survey ($)	0.13	0.55	0.07	1.15	22.07	0.05

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.

We focus on variations in cost by survey respondent type, since Facebook ads are deployed using a bidding system in which hard-to-reach populations are more expensive for advertisers to target.

In Indonesia, the data for the 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.

S5.1 Variation in costs by strata

One artefact of Facebook's bidding-based advertising system is that some strata are substantially more costly. 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 and 0.7-6% in Kenya. 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 the spread was much larger, from $0.05 - 22.07 (despite the fact that we periodically switched off high-cost strata).

Generally, we found that targeting in Mexico was cheaper and more successful than in Kenya. This likely resulted from two constraints: (1) the lower levels of Facebook usage in Kenya, and (2) the fact that our Facebook targeting criteria in Kenya were more specific with respect to geography, thereby making it more difficult to fill survey strata. Although we incentivized respondents in Kenya (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.

S5.2 Costs by round of data collection

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 S4 and S5, 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 over-sample 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.

	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

Table S4: Reach of surveys in Mexico and Kenya, split into main sample and low-education (in Mexico and Kenya) / high age (in Kenya) oversamples

	Initial			Oversample			Initial			Oversample		
	Median	Max	Min	Median	Max	Min	Median	Max	Min	Median	Max	Min
Click through rate (%)	6.64	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 ($)	0.02	0.05	0.01	0.07	0.18	0.02	0.18	0.47	0.02	0.09	0.33	0.04
Cost per completed survey ($)	0.07	0.31	0.02	0.58	4.50	0.08	1.19	22.07	0.05	0.78	8.72	0.22

Notes: The click through rate and the completion rate are defined with respect to the number of impressions.

Table S5: Cost of Facebook targeting in Mexico with and without low-education oversample.

S5.3 Variation in costs for specific populations of interest

Above, we discussed the fact that survey costs can vary dramatically by strata and by the nature of a given survey round. This translates into practical tradeoffs and might create disincentives to target certain populations, and these disincentives can be substantial. Concretely speaking, the difference between the lowest ($0.05) and highest ($22.07) costs we observed per survey for different strata in Kenya would imply that it is cheaper to obtain surveys from 441 men in Nairobi than it is to reach a single female respondent at a specific part of the border between Kenya's Coast and North-Eastern provinces. In Mexico, the highest cost cell was $4.50 for women above the age of 60 living in rural Southern Mexico. The lowest cost cell was $0.02, by contrast, for men between the ages of 50 and 59 living in rural Northern Mexico.

In Table S6, we present the average cost of targeting different respondent types. For example, to compute the average cost per male respondent, we averaged the cluster-specific cost per response for all clusters containing men. Using this approach, we estimate that Kenyan women are over 4 times more costly to target than Kenyan men, whereas rural respondents are over 5 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 municipalies are not as starkly different as in Kenya.

S6 Weighting

For both 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. 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),

	Mexico FB	Kenya FB	Indonesia FB
Overall	$0.16	$1.06	$
Men	$0.15	$0.45	$
Women	$0.18	$1.89	$
Urban	$0.12	$0.28*	$
Rural	$0.21	$1.53*	$

Note: Afrobarometer sites are classified as rural, urban, or both. When aggregating to k-means clusters, we defined "rural" cluster centroids as centroids for which there were more rural than urban Afrobarometer sites (n=34), and likewise for "urban" k-means cluster centroids (n=10). The calculations ignore Afrobarometer sites which are classified as "both." They also ignore k-means cluster centroids for which an equal number of clusters were rural and urban (n=6). In the Mexico case, we defined as "rural " those respondents 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).

Table S6: Cost per response for the given respondent type, averaged over all clusters with those respondent types

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 both 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 J., Michele F. Margolis, and Michael W. Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58 (3): 739–753. ISSN: 1540-5907. doi:10.1111/ajps.12081.

Berinsky, Adam J., Michele F. Margolis, Michael W. Sances, and Christopher Warshaw. 2019. "Using screeners to measure respondent attention on self-administered surveys: Which items and how many?" *Political Science Research and Methods:* 1–8. ISSN: 2049-8470, 2049-8489. doi: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.

Mercer, Andrew, Arnold Lau, and Courtney Kennedy. 2018. *For weighting online opt-in samples, what matters most?* https://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/.

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