1

Modern Polling

Challenges and Opportunities

Polling has had a rough ride recently. In 2016, pundits confidently predicted Hillary Clinton would be elected president, bolstered by survey evidence suggesting she led nationally and in the critical swing states needed to win the Electoral College. Donald Trump's victory shocked the world and many blamed polls for misleading them. In 2020, Donald Trump didn't win, but he came much closer than polls suggested. Almost every poll showed Joe Biden leading nationally and in crucial swing states – often by healthy margins. Biden did, in fact, prevail in the popular vote and most swing states, but not easily. Had 40,000 more voters in three states voted for Donald Trump, Trump would have shocked the world again. The shortcomings of the 2020 polls were particularly unnerving because pollsters tried to learn from the mistakes of 2016, only to produce the least accurate polls in 40 years (Clinton et al. 2021).

Problematic polling is not uniquely American. In October 2022, polls suggested that former President Luiz Inacio Lula da Silva led right-wing populist President Jair Bolsonaro by 14 percentage points in the first round of Brazil's presidential election. On election day, Bolsonaro lost by only 5 percentage points. Populists also outperformed polls in the 2016 Brexit vote in the United Kingdom. The polls showed the pro-Europe "remain" and populist "leave" sides to be very close, with the remain side generally ahead. On election day, the leave side won by 4 percentage points (Hanretty 2016).

Polling misfires like these have soured many people on the field. "If 2016 didn't prove it, 2020 certainly did: The polling industry is in crisis" (Roberts 2020). An editorial in the *Wall Street Journal* didn't mince words: "It's all garbage . . . we now have convincing evidence that polling fails to measure accurately that which it seeks to measure, and pollsters have no idea why" (Stemberg 2020). For Republican pollster Frank Luntz, "the political polling profession is done" (Concha 2020). Perhaps polls need to go the way of phone books, video stores, and manual transmission cars. Parts of the public wouldn't miss them: Polls show (ironically) that a majority of Americans – and a super majority of conservatives – do not believe polls (Sheffield 2018).

https://doi.org/10.1017/9781108697798.002 Published online by Cambridge University Press

Polling experts tend to be more sanguine. They note that polls remain generally accurate, across time and countries (Jennings and Wlezian 2018; Morris 2022a; Silver 2021a). Polls performed especially well in the 2018 and 2022 congressional elections (Silver 2018; Morris 2022b). These experts worry less about bad predictions and more about the decline of random sampling.

Random sampling theory has traditionally anchored survey work, justifying how characteristics of relatively small groups of people will converge to the characteristics of a population as long as the small group is a random sample of the large one. But random sampling is for all intents and purposes dead. A truly random sample requires that people are randomly contacted and that they *all* respond. In 2018, the response rate for the *New York Times*/Sienna College polls was less than 2 percent. The raw numbers are staggering: They made 301,921 calls to get 5,612 responses. In some districts, only 1 percent responded; in Michigan's 8th congressional district, the pollsters made 40,230 calls to get 447 responses. They had particular difficulty reaching young people who responded only 0.5 percent of the time.[1]

Response rates have continued to deteriorate since 2018 (Cohn 2022a). They are now so low that pollsters doubt the future of traditional polling efforts: "If you were employed as one of our interviewers at a call center, you would have to dial numbers for two hours to get a single completed interview … this is getting pretty close to 'death of telephone polling' numbers" (Cohn 2022a).

Many pollsters no longer even attempt random sampling. Typically, these nonprobability pollsters harvest contact information from internet ads and then use complex algorithms to decide whom to contact and how to weight their samples. Much of the energy in contemporary polling is devoted to refining these nonprobability methods. While nonprobability polls are often accurate, they require stronger assumptions than random sampling.

Polling has therefore reached a critical moment. Polls have failed us in high-profile elections and, at best, have a weak basis in the random sampling theory that historically justified polling. The goal of this book is to help us understand this moment and plot a course for where to go next.

To understand this moment, I interpret the history and current practices of polling in terms of two sources of polling bias. One arises when survey samples differ from the population with respect to demographic and other easily measurable characteristics. The other source of bias arises when survey samples differ from the population with respect to less easily measurable characteristics, such as social trust or even the concept being measured in the polls. This second source of bias is particularly tricky because it confounds the standard tools of contemporary survey analysis. For example, if the ideological or psychological attributes that attract some people to Trump also lower

[1] These response rates may be lower than for other polls because they called more people who were unlikely to answer (Cohn 2018).

their propensity to respond to polls, demographic adjustments will not offset artificially low support for Trump among survey respondents.

Traditionally, random sampling saved us from both sources of bias because as a random sample gets larger, *all* characteristics of the sample – observed or not – converge to the distribution of characteristics in the population. Now that random samples have become rare – and possibly extinct (due to nonresponse and increasing use of nonprobability samples) – we need a new framework that explains error in the nonrandom samples that modern polls produce. I focus in this book on Meng's (2019) framework because it elegantly characterizes polling error in all surveys, randomly sampled or not. Meng's paradigm doesn't contradict random sampling theory; random sampling is a special case and continues to be an attractive ideal. But Meng's more general framework helps us focus on *all* the ways that modern surveys can go wrong.

Some implications of Meng's framework are familiar. Error in polls is inversely proportional to sample size, just as in random sampling. Some implications are less familiar. Bias of the second type – bias that arises when respondent's willingness to respond depends on the variable being measured by the survey – creates errors that are directly proportional to the size of population being measured. For survey researchers trained in random sampling, linking sampling error to population size (as opposed to sample size) verges on heretical. And yet this result is easy to explain in Meng's framework.

Connecting sampling error to population size means that small amounts of nonresponse bias can metastasize into large error in nonrandom samples, making it imperative that we do everything we can to prevent or adjust for all sources of bias. Conventional methods handle one, but not both, sources, creating a dangerous blind spot. This book will show that handling both sources of error requires new methods and new types of data. In explaining how and why these new approaches work, this book maps out next steps to help us design surveys to address the full range of survey bias in a post-random sampling world.

The book is based on three premises. First, polls are important. We simply cannot understand politics, society, and the economy without them. Social scientists and survey methodologists need to do whatever they can to get them right. Second, getting polls right is not easy, however. Extrapolating from the responses of the increasingly rare and potentially atypical people who respond to modern polls is daunting, making polling vastly more difficult than when random sampling was viable. Therefore, we need to think deeply about all the ways that polls can go wrong in this new environment, not just the ways that are easy to fix. Third, to say that getting polling right is difficult is not to say that it is impossible. While some may believe that polls will never be able to correct for biases caused by unmeasured factors, this book shows that new theories, data, and methods can substantially improve our ability to account for this kind of bias as well.

This chapter sets the stage for the book. Section 1.1 emphasizes how important accurate polling is for policymakers, researchers, and citizens. Section 1.2 identifies two fundamental sources of bias in nonrandom samples. Section 1.3 explains how current approaches to polling address only one of these sources. Section 1.4 previews the tools presented in this book that enable us to address both challenges. Section 1.5 provides an overview of the chapters of the book, and Section 1.6 describes whom this book is for.

1.1 WE NEED GOOD POLLS

Polls are essential to a well-functioning modern society. They provide an important window into the beliefs, desires, and realities of everyone in society. In a world without polls, we could still learn about these things with in-depth qualitative research (e.g., Cramer 2016), but it is really hard to keep such information steadily flowing in a way that is comparable, generalizable, and free from subjectivity.

Polls are especially useful to burst epistemic bubbles. An old chestnut in polling circles tells of a New York socialite who, after Richard Nixon's landslide win in 1972, stated that she didn't understand how Nixon could have won because she did not know anyone who voted for Nixon. The story turns out to be apocryphal, but circulates because it is very funny – and a little true. We all are tempted to think that everyone sees the world as we do. Polls make that harder because they measure the views of people who may experience the world dramatically differently.

Some believe that flaws in modern polling corrode their ability to inform political leaders about public preferences. Matt Yglesias (2022) worries about survey samples dominated by high-intensity voters, writing

> the answers-polls electorate is to the left of the votes-in-elections electorate, so Democrats keep getting a skewed read of the landscape and mis-calibrating their own races. The actual country is simply less-educated, lower in social trust and openness to experience, and more right-wing than the country that shows up in surveys. Democrats can win those voters; they just need to realize the practical necessity of doing it.

A similar dynamic may occur for Republicans, especially Republican politicians in gerrymandered districts that are so strongly Republican that they worry more about primary challenges than general election races.

Polling's loss of credibility played out in interesting ways in the 2022 US congressional campaign. Preelection polls from traditional probability-based pollsters tended to show good results for Democrats. Nonprobability polling tended to favor Republicans. Having been burned in 2016 and 2020, many in the media assumed the Republican-leaning polls were more accurate. Their articles anticipated a bad night for Democrats: "Democrats' midterm hopes fade: 'We peaked a little early'" (*Politico*) or "Top Democrats Question Their Party's Strategy as Midterm Worries Grow" (*New York Times*) or even "Red Tsunami Watch" (*Axios*).

https://doi.org/10.1017/9781108697798.002 Published online by Cambridge University Press

Some Republican insiders were particularly confident, in part because they believed that Republican support was strong among nonrespondents. Benjamin Wallace-Wells (2022) summarized what he was hearing from his Republican sources:[2]

[T]he polls were likely under representing certain segments of the electorate. In recent years, more educated voters, especially white women, have moved to the Democrats, and less educated ones, of all races and especially men, toward the Republicans. When it comes to polling, these shifts have created an imbalance, in which one of the most visible groups in politics, and one especially energized by the *Dobbs* decision, had shifted toward Democrats, and one of the least visible had shifted toward Republicans. "The fastest-moving portion of the electorate is Hispanic men, and the second-fastest-moving portion of the electorate is Black men," the Republican consultant told me. You want to get them on the phone? "Good f—ing luck."

Issue polling may be even harder, because unlike campaign polling, the true "answer" about public preferences is seldom revealed. Surveys during the Covid pandemic showed support for masking in response to Covid. What if people opposed to masking were less likely to respond? Polling often shows Republicans support tax cuts. What if the Republicans who respond to polls are more likely to support party initiatives than those who do not respond? After the police killing of George Floyd in 2020, surveys suggested that attitudes of white people shifted on race. Was this real movement? Or was it simply new patterns in nonresponse? The implications for understanding how society was responding to a crisis in race relations are profound (Graham 2020).[3]

Polling problems may even undermine democracy. Some Republicans feel that polls are another elite institution biased against them. James Baker, the former US Secretary of State and manager of five Republican presidential campaigns, took to the pages of the *Wall Street Journal* immediately after the 2020 election to argue that

It would be funny if it weren't a sad reality that American democracy is being undermined by bad polling.... Polls that repeatedly favor one side create false expectations that adversely influence the actions of both sides. The favored side becomes overconfident and suffers when the results on Election Day don't meet expectations. And the

[2] The article was aptly, but not presciently, titled "Why Republican Insiders Think the G.O.P. Is Poised for a Blowout: The Consensus among Pollsters and Consultants Is This Tuesday's Election Will Be a 'Bloodbath' for the Democratic Party."

[3] Some researchers defend issue polling because the conventional polls match larger federal polls on many demographic and lifestyle benchmarks (Kennedy, Mercer, Hatley, and Lau 2022). Finding one set of polls matches another is informative but does not definitely show that issue polling is accurate. Even in this analysis, the conventional poll differed substantially from federal polling benchmarks in several respects. Respondents to the conventional poll were 21 percentage points more likely to say they had a retirement account, 11 percentage points more likely to say they voted, and 8 percentage points less likely to say they received food stamps.

https://doi.org/10.1017/9781108697798.002 Published online by Cambridge University Press

disfavored side is disadvantaged in both fundraising and voter turnout by the appearance that the outcome is foreordained. ... It's little wonder that many Americans — especially Republicans — voice complaints when newspapers and networks misinform them on the state of an election. ... Being branded "rank propagandists," as one knowledgeable observer recently called pollsters, cannot be good for business. (Baker 2020)

Former President Trump was not above trying to take advantage of the state of affairs, claiming without evidence that his wild claims about the 2020 election had widespread support from the American public. "He might be wrong, but without reliable polls, who's to say otherwise?" (Graham 2020).[4]

1.2 WHY POLLING IS DIFFICULT

The problem of modern polling is not a problem with random sampling. Random sampling is a powerful tool for ascertaining the views of many from the responses of a few. It works because random samples will look more and more like the population from which they were drawn with regard to *all* features – measured or not – as the size of sample increases. But a sample is only random in this way if *everyone* who is randomly contacted responds. This has never literally been the norm, but the polling world has generally been comfortable using samples in which most people respond to approximate random sampling, especially when they reweight survey samples to match known population demographics.

The problem of modern polling is a *lack* of random sampling. Today's abysmally low response rates make it impossible to pretend that samples are random. The *New York Times* contacts a random group of people, but when only one or two in 100 respond, they hear from a nonrandom subset that is unlikely to constitute a random sample.[5] The people who responded differ from the target population if for no other reason than they chose to do something that most people do not do: Respond to a poll. Survey respondents are in a select group of the top 1 percent of people most willing to respond to a poll. For comparison, the top 1 percent of US households earn more than $700,000 per year (Mishel and Kandra 2020). The tallest 1 percent of American men are over 6 foot 4 inches tall. The tallest 1 percent of women are over 5 foot 10 inches tall.

[4] Trump's belief in polls is as Trumpian as one might guess. At a 2018 campaign rally, he stated "I believe in polls — only the ones that have us up.... Other than that, they're the fake news polls" (Sheffield 2018).

[5] It's even hard these days to identify a target population from which to sample. In the past, most households had a landline phone linked to a specific geographic locale. Now people have complicated mixes of cell and landline phones, many of which are untethered to any geographic region, meaning that there is neither a white pages-type directory of people living in a region nor a guarantee that a randomly dialed cell phone will be answered by someone living where we think they live.

https://doi.org/10.1017/9781108697798.002 Published online by Cambridge University Press

If we're going to extrapolate from the nonrandom people who answer surveys to the general population, we will need to adjust for their distinctive character. To do this, pollsters identify two ways that respondents and nonrespondents can differ. The first is with respect to variables that we can observe, often variables such as gender, age, race, income, and education. Such differences are common. In US surveys, respondents are often older, whiter, richer, and more educated than the population at large. This type of nonresponse is concerning, but relatively easy to fix. Suppose there are too few young people in a survey sample. Perhaps we observed 100 young people even though Census data indicated that a random sample should have yielded a sample with 200 young people. We can simply double count the responses of the young people we have, and our adjusted sample will have the correct proportion of young people.[6] The trick here is filling in for young people who didn't respond with the answers of the young people who did.

In academic parlance, nonresponse associated with observed characteristics such as age and education is called "ignorable nonresponse." The reason this nonresponse is called ignorable is that by adjusting the data with weighting (or a related technique), we render this kind of nonresponse harmless and, hence, ignorable.

We use this terminology because it is standard in the literature. It's not great, though. One might think, reasonably enough, that ignorable nonresponse can be ignored. That's not true. Pollsters do not ignore ignorable nonresponse. They diligently weight their way around it. The label is intended to capture the idea that *if* we do something (like weighting), *then* we can ignore it. A better description of ignorable nonresponse is "conditionally independent nonresponse," because ignorable nonresponse is independent of survey opinions once variables observed in the sample and population are accounted for. This means that while young people may have different views than old people, the young people who responded have on average the same views as the young people who did not and the old people who responded have on average the same views as the old people who did not.

The second way that nonrespondents can differ from respondents is more subtle. Nonrespondents may differ with regard to some characteristic related to the quantity we are trying to measure. For example, suppose that we are trying to measure political views such as ideology and that respondents are more liberal than nonrespondents, perhaps because liberals may be more willing to trust pollsters. Obviously, this creates problems because our sample will be too liberal. And we cannot use weights because we don't know how many people are liberal in the whole population – that's what we were trying to measure with the survey!

[6] Weighting is second nature for experienced pollsters. For people new to the idea, it can seem like cheating. I explain weighting in detail in Chapter 3.

In academic terms, this is called "nonignorable" nonresponse because even after adjusting our sample data based on observable characteristics, this nonresponse will cause bias and hence cannot be ignored. At its root, nonignorable nonresponse arises when people's willingness to respond is related to their opinions even after accounting for demographics. This can happen in two ways. First, a factor might explain both response propensity and the variable being measured in the survey. Many suspect, for example, that socially trusting people are more willing to respond to polls and are more likely to be liberal. Second, the outcome variable itself could affect response propensity. This can happen if someone is more likely to respond to a customer satisfaction survey after having a bad experience.

I also use this label because it is standard in the literature on polling, but I do so grudgingly. Nonignorable nonresponse is not so much a label as an instruction. Do. Not. Ignore. The point is that we should not ignore this kind of nonresponse because it will bias our estimates even if we weight our data. If you prefer using a more informative label for nonignorable nonresponse, consider "outcome-related nonresponse," which more directly indicates that nonignorable nonresponse occurs when the survey outcome relates to nonresponse.[7]

There are many contexts in which nonignorable nonresponse is a legitimate concern. In political polling, for example, nonresponse appears nonignorable for questions about intention to vote (Lohr 1999, 256). This is not surprising because habitual voters tend to be more interested in politics and more willing to engage in the rituals of citizenship, ranging from talking to pollsters to turning out to vote. Voters are more likely to respond to polls even when we account for demographics, meaning that we cannot weight our way back to a representative sample because in every possible demographic group we'll expect to see higher turnout among respondents than among the population at large.

The following are examples in a wide range of contexts.

- In survey research, the opinions or characteristics being measured may directly influence people's willingness to respond, especially on sensitive topics. Whether a person owns a gun may directly affect whether they respond to a poll about gun ownership (Urbatsch 2018). Such dynamics may occur on surveys about race, sexuality, sexual harassment (Cantor et al. 2020), and other controversial topics.

[7] Some researchers use a different, but related, set of terms in which data produced by an ignorable nonresponse process is also characterized as "missing at random" (see, e.g., Lohr 1999, 265). A strong form of ignorable nonresponse occurs when data are "missing completely at random." In this case, there is no covariate that explains the missingness. This occurs, for example, when we have 100 percent response rates from a randomly selected sample of the population. When data are missing at random (in this technical sense), nonresponse is random once we have accounted for observed characteristics such as age, race, and gender.

- Marketing researchers interested in online reviews face likely nonignorable nonresponse because such feedback skews strongly positive (Schoenmueller, Netzer, and Stahl 2020). There is also considerable evidence that respondents to nonprobability internet surveys are more active on the internet, more likely to be early adopters, less traditional and more environmentally concerned than the population at large, attributes that seldom are included in weights but could affect opinions about products (Gittelman et al. 2015).
- In economics, researchers worry that the least and most well-off are the most reluctant to reveal their income on surveys (Bollinger and Hirsch 2013).
- In communication studies, extrapolating from any social media data to a population is extremely challenging. Not only may social media users be younger, wealthier, and more technically skilled (Hargittai 2020), their willingness to use social media may reflect distinctive personality traits and worldviews.
- In public health, the propensity of people to respond may be related to the behaviors and outcomes being surveyed. A health study in the United Kingdom tracked down nonrespondents and found that they were 59 percent more likely to be in poor health than respondents (Peytchev 2013, 94). Nonignorable nonresponse is a concern on vaccination surveys because people who get vaccinated may be more likely to respond, as we see later.

Ironically, most survey work ignores nonignorable nonresponse. Virtually every survey analysis in both applied and academic settings uses only weighting or another tool that counters only ignorable nonresponse. Part of this puzzling state of affairs is explained by the evolution of modern survey research that was built on random sampling foundations. When using random sampling, pollsters did not have to worry about the types of nonresponse because random sampling countered both ignorable and nonignorable nonresponse. As response rates declined, pollsters retained the random sampling justification – producing margins of error based on random sampling, for example – and used weights to tweak their nonrandom samples to make them look like they were actually random samples. Not having moved to more general theory most pollsters have not felt an urgency to move beyond weights and similar tools that address only ignorable nonresponse.

Another reason pollsters ignore nonignorable nonresponse is that many believe that nothing can be done. One reason for this belief seems logical: If the problem is based on factors we do not measure, surely we cannot do anything about it. We shall see that this logic is incorrect. Another reason is that few pollsters are trained to address nonignorable nonresponse. Even though tools that counter nonignorable nonresponse have been around for decades (e.g., the Heckman selection model that I discuss in Chapter 8), they are seldom emphasized in the training of survey researchers. The low profile of these tools may have arisen in part because they often proved unreliable in practice. I describe in the second half of this book the considerable methodological progress that has been made beyond the traditional tools. More importantly, perhaps, I emphasize that even the older tools can work well if we design our surveys to produce the kind of data they require.

1.3 THE LIMITS OF CONVENTIONAL PRACTICE

What should we do about the nonresponse bias that can arise in modern nonrandom samples? The appropriate methods depend on the nature of the nonresponse. If nonresponse is ignorable, the conventional polling toolkit provides excellent options. That is, conventional tools are good at adjusting for nonresponse when the differences between respondents and the full population can be explained by variables that (1) we observe for respondents and (2) for which population distributions are known (Mellon and Prosser 2017, 772). These tools adjust the sample data such that the distribution of these observable variables in the sample matches the population distribution. Weighting is the most widely used tool; Chapter 4 discusses other tools that perform the same task based on the same assumptions, tools such as quota sampling.

Conventional polling practice has less to say about how to counter nonignorable nonresponse. As discussed earlier, the most common approach to nonignorable nonresponse is to ignore it. Ignoring the problem is defeatist; the goal of social science should be to identify and fix problems. Other options in the standard toolkit include eliminating nonresponse, weighting and increasing sample size (Lohr 1999, 256). None of these options is appealing. Eliminating nonresponse, for example, is almost never feasible.

Weighting is the most common way to attempt to deal with nonignorable nonresponse. Does weighting as conventionally practiced at least help fight nonignorable nonresponse? Sadly, no. Weighting can even exacerbate bias. As we alluded to above, weighting only works if respondents are random samples of their demographic groups (Groves et al. 2009, 350). We've seen that if a survey sample doesn't have enough young people, a pollster using weights will use the responses of the young people who did respond to speak for those who did not. *If* the young people in the sample are representative of all people in the population, weighting is great. But what if nonresponse is nonignorable, meaning that the young people who answer polls are not a random sample of their peers? In this case, weighting does not correct for nonresponse bias. In fact, if the young respondents are particularly unusual, up-weighting them may make things worse because weighting attributes beliefs to the nonrespondents that do not reflect their opinions (Agiesta 2021). Ken Goldstein (2016) summarizes the problem: "Usually we assume the problem is that group *X* is too small, but the actual problem may be that group *X* is too weird." Placing *extra* weight on members of these group *X* weirdos will make our estimates worse, not better.

Within the weighting paradigm, one counters nonignorable nonresponse by finding weights that absorb the heretofore unmeasured sources of bias. If someone can find a way to measure the characteristic at the root of the nonignorable nonresponse in *both* the sample and population, they could add the variable to their weighting protocols and defang the nonignorable nonresponse. For example, if one believes that low levels of trust are associated with both nonresponse and support for populist politicians, one could theoretically measure trust among survey respondents and among the population at large and use it as a weighting variable. The limits of this strategy are clear, however: The only way to measure trust in the population is with a survey, which itself would be subject to nonresponse bias. Such circularity makes it hard to believe that finding new variables to include in weights will definitively solve nonignorable nonresponse.

In recent years, we have been awash in big data and it is tempting to think that these large data sources offer a way to counter nonignorable nonresponse. This doesn't work either. Not only do large sample sizes not address nonignorable nonresponse, they may make matters worse (Meng 2018). Bradley et al. (2021) provide a vivid example related to vaccination rates. The specter of nonignorable nonresponse looms over vaccination studies because it may be easier to contact and get response from people who do socially acceptable things like getting vaccinated.

Bradley and colleagues compared the following three surveys about Covid vaccination to US government baselines.

1. A Facebook survey produced more than 4.5 million responses across multiple waves over time. This survey weighted on age and gender but not education or race/ethnicity. Because the survey was conducted on Facebook, it was limited to Facebook users.
2. A Census Household Pulse survey produced more than 600,000 responses across multiple waves. This survey was an experiment by Census designed to rapidly measure pandemic-related behavior. They randomly sampled people for whom they had cell phone or email contact information, a population that included approximately 81 percent of US households. Response rates were under 10 percent (Peterson, Toribio, Farber, and Hornick 2021). This survey weighted on age, gender, education, and race/ethnicity.
3. An Axios-Ipsos survey of around 1,000 respondents per wave based on samples from all addresses in the United States. The survey firm provided internet access to respondents who lacked internet access (accounting for about 1 percent of the final sample). This survey weighted on age, gender, education, and race/ethnicity.

What should we expect from these surveys? A big data enthusiast would love the large samples of the first two surveys. A technically oriented pollster might be attracted to the Axios-Ipsos survey because they had the strongest argument that they were sampling from the whole country, not just Facebook users or people for whom the Census Bureau had contact information.

The Axios-Ipsos survey, small sample size notwithstanding, consistently tracked best with the government baseline with the reference data from the CDC in or nearly in the confidence interval produced by the Axios-Ipsos survey for the January through May 2021 period that was analyzed. Despite their much larger sample sizes, the confidence intervals of the Facebook and Census surveys did not contain the CDC reference estimates, meaning they were inaccurate. And the problem grew over time. By March 2021, the Census measure of vaccination take-up was 9 percentage points higher than the CDC measure and the Facebook measure was 16 percentage points higher than the CDC measure. These differences were even larger by May 2021.

Why did the smaller sample size Axios-Ipsos survey perform best? Bradley et al. (2021) argue that survey methods are more important than sample size. That is, the smaller-sized survey performed best because it was less likely to suffer from nonignorable nonresponse bias than the big data surveys (Meng 2018).

Given the dearth of tools in the conventional toolkit that deals with nonignorable nonresponse, some advocate simply coming clean: Polls will be imperfect and the public needs to accept this (Leonhardt 2020). In this approach, pollsters should emphasize that errors are inevitable, especially today because there are so many ways for the nonrandom samples produced by modern polling to go sideways. In other words, polls will sometimes be wrong. Get over it.

1.4 THE ROAD AHEAD: A PREVIEW

As social scientists, we should resist this polls-will-be-polls attitude. I advocate another option: Constructing a polling paradigm that accounts for ignorable nonresponse (the nonresponse that weighting can fix) and nonignorable nonresponse (the nonresponse that weighting cannot fix). This paradigm can broaden our intuitions about when polling does and does not work. It also helps us focus attention on designing and analyzing surveys in ways that are less vulnerable to all types of nonresponse that can arise in modern nonrandom samples.

There are, broadly, three approaches to dealing with nonignorable nonresponse. The first involves documenting how vulnerable any given survey is to nonignorable nonresponse (Hartman and Huang 2023). Suppose that we are concerned that our survey has too many politically interested people. We cannot weight for political interest because we don't measure political interest among nonrespondents. We can, however, posit different possibilities (e.g.,

https://doi.org/10.1017/9781108697798.002 Published online by Cambridge University Press

"what if 30 percent of people are interested in politics? Or 20 percent?") and for each possibility generate weights.

Such sensitivity analysis enables us to estimate how vulnerable a survey is to nonignorable nonresponse. If our conclusions change dramatically over the range of posited population averages, the results are indeed vulnerable to nonignorable nonresponse and we need to inform our readers accordingly. Even better, we should implement surveys following the ideas outlined in this book. If the results do not vary much across the posited population values, we can move our survey out from under the cloud of suspicion with regard to at least that source of nonignorable nonresponse. Chapter 7 discusses sensitivity analysis.

A second approach to dealing with nonignorable nonresponse involves using existing data in new ways. The 2020 American National Election Study (ANES) provides an interesting case. The ANES is a long running and widely respected academic poll. In the ANES preelection survey data, Biden led Trump by 11.8 percentage points (53.4 to 41.6), a lead that exceeded Biden's actual 4.4 percentage point margin victory in the popular vote.[8]

Figure 1.1 displays ANES results by political interest. The top panel shows support for Biden organized by interest in politics. Among respondents very interested in politics, Biden support was 61.9 percent. Support for Biden among somewhat interested people was 56.5 percent, while support for Biden among respondents who indicated they were not very or not at all interested in politics was 51.1 and 49.0 percent, respectively. In short, support for Biden among the most interested respondents was 12.9 percentage points higher than among those least interested in politics.

The bottom panel presents a similar breakdown for feeling thermometers, answers people give when asked to rate a politician on a 0 to 100 scale. Among people very interested in politics, the feeling thermometers for Biden averaged 19.3 points higher than the feeling thermometers for Trump. Among people somewhat interested in politics the gap was 9.3, while among people who said they were not very or not at all interested in politics, average feeling thermometer differences between Biden and Trump were only 2.8 and 0.7, respectively.[9]

If we believe that people who are more interested in politics are more likely to answer a poll about politics – which hardly seems farfetched – then it seems natural to worry that the ANES suffered from nonresponse bias. The response

[8] Using ANES's weights actually increased Biden's margin to 12.6 percentage points. The ANES was not alone in suggesting Biden's lead was large. The final preelection FiveThirtyEight.com margin based on many polls indicated that Biden led by 8.3 percentage points. Across respected large-scale polls from the ANES, the Cooperative Election Study, Nationscape and Pew, the average Biden lead was 17.9 percentage points unweighted and 14.7 percentage points when weighted (Jacobson 2022, 9).

[9] Controlling for demographics does not change the patterns reported in the figure. I'm grateful to Leonie Huddy for suggesting this example.

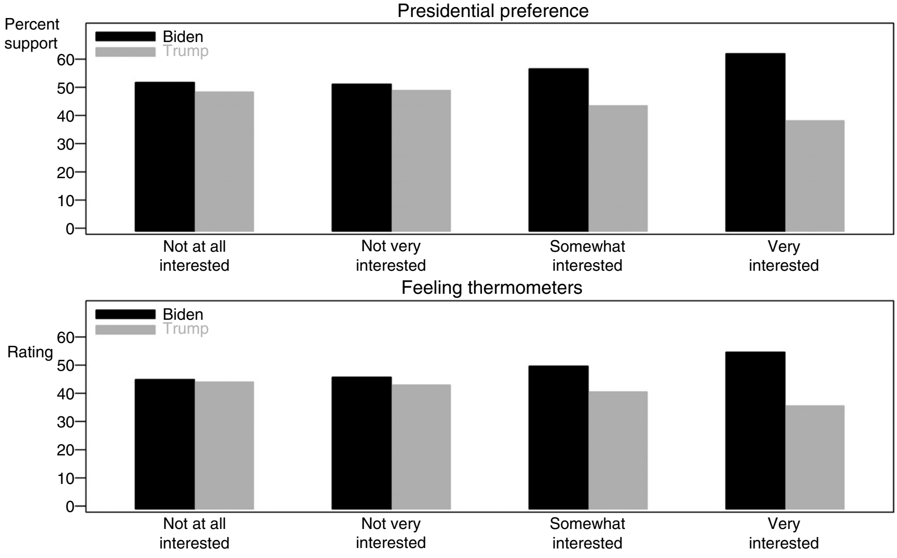

FIGURE 1.1 Support for Biden and feeling thermometer difference between Biden and Trump

rate for the ANES poll was 36.7 percent, so a lot of people didn't respond. If those nonrespondents were less interested in politics, then we should probably extrapolate lower support among nonrespondents than among respondents. Doing so would adjust Biden's lead downward toward the actual result.

While conceptually simple, no pollster that I know of made an adjustment based on interest in politics. Why? As we have already discussed, using nondemographic variables in weighting is difficult – if not impossible – because weighting requires us to know the distribution of weighting variables in the population. Because we don't know the actual distribution of political interest in the population, we don't weight for it, leaving us to disregard this seemingly obvious possible source of nonignorable nonresponse.

It need not be this way. Part III of this book describes tools that measure and correct for nonignorable nonresponse using such data. Peress (2010) did this to model turnout in the 1980s. As is often the case, surveys at that time overestimated turnout: Even though only 50 percent of adults turned out to vote at that time, 70 percent of ANES respondents voted.[10] Weights took a chunk out of the bias – turnout in the weighted data went down to around 60 percent – but did not eliminate the problem. Using the approach discussed in Chapter 9, Peress incorporated information about response interest, information that is akin to the political interest variable discussed plotted above and

[10] The ANES validated whether people voted so this was not simply a case of people saying they voted when they didn't. See Jackman and Spahn (2019) for excellent deep dive into this issue.

https://doi.org/10.1017/9781108697798.002 Published online by Cambridge University Press

was able to bring estimates to within 1 percent of actual turnout in 1980 and 1988 and within 2 percent in 1984.

A third type of tool for tackling nonignorable nonresponse involves survey design. The importance of survey design is deeply embedded in polling DNA; after all, in random sampling the design of the survey is much more important than the number of responses. The centrality of design persists in the new paradigm as the goal is to create data that make it easier to diagnose and potentially correct for nonignorable nonresponse.

To explain how this works, let's start with a classroom scenario. Suppose that we are interested in knowing how many students know the answer to a test question. We'll assume there are two classes, each with 50 students, half of whom know the right answer. The teacher in the first classroom randomly selects five students and gives them the quiz. The teacher in the second classroom asks for volunteers and has the first five students who raise their hands do the quiz.

Will the quiz scores be the same in both classes? Probably not. In the first class, we should expect that 50 percent of the students will answer correctly because they have been randomly selected. In the second class, it seems likely that the students who volunteer will be more likely to get the correct answer.

Now suppose the teachers increase the number of students taking the quiz to 15, which is 30 percent of the class. In the first class, we should still expect 50 percent correct answers. In the second class, the scores may go down. We're still getting volunteers and volunteers are more likely than nonvolunteers to know the answer, but now we're going deeper in the pool. So even as the percent correct would likely exceed 50 percent, it would not be as high as when only five students took the quiz. If we keep increasing the number of students who take the quiz, eventually everyone in both classes will take the quiz and both classes will generate 50 percent correct answers.

The nonresponse in the first class is ignorable because students are randomly selected. In other words, in that first class there could be no systematic relationship between who responded and whether they knew the answer. In that class the percent correct was 50 percent in expectation whether 5, 15, or all 50 students took the quiz. In the second class, nonresponse is nonignorable because whether a student knew the answer influenced whether they volunteered. In that class, the average quiz score declined as the number of quiz-takers went from 5 to 15 to 50.

Figure 1.2 illustrates the scenario. On the y-axis is the percent correct. On the x-axis is the response rate. The line on top shows that the expected percent correct declined as the response rate increased for the voluntary classroom in which nonresponse was nonignorable. When 10 percent of the students volunteered, the percent correct was clearly higher than when 30 percent of the class volunteered, which was higher than when everyone answered. The flat line at 50 percent shows the expected percent correct for the randomly selected class. It was always the same regardless of response rate.

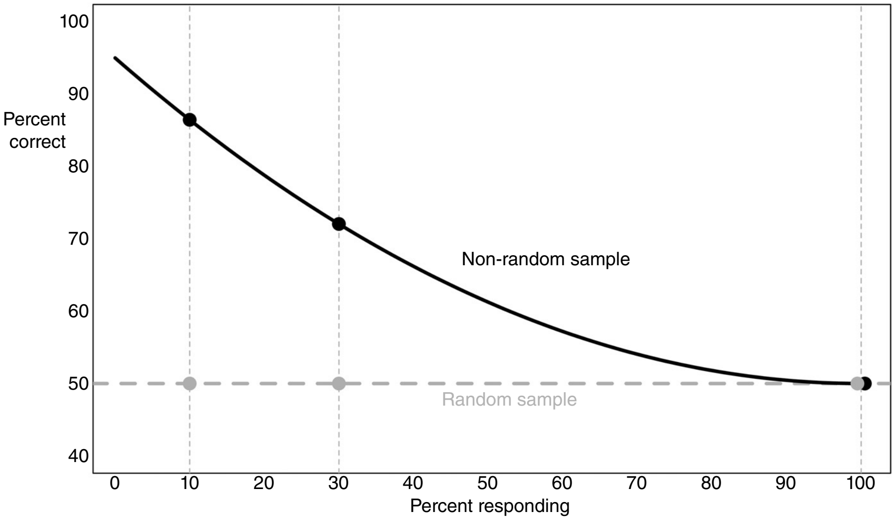

FIGURE 1.2 Relationship between response rates and quiz performance for two types of nonresponse

Now we can extend the logic to polling. Suppose we field two surveys simultaneously, one with a low response rate and one with a high response rate. Cohn (2022b) reports one example in which the *New York Times*/Siena pollsters fielded a conventional phone survey with typical low response rate around 1 percent and another high response survey in which respondents were paid to respond (and did so at a 30 percent rate). If the percent "correct" – which in this case could be support for a given political candidate – comes back different in the two arms of the survey, which line in Figure 1.2 would that be consistent with? If the percent supporting the candidate were the same, which line in the figure would that be consistent with?

Following our logic so far, different results across survey arms would accord with nonignorable nonresponse and similar results across surveys would accord with ignorable nonresponse. This, then, is the core logic for pulling out evidence of nonignorable nonresponse. There is work to be done, of course, in moving from the intuition to statistical models and survey design, work we do in Chapters 10 through 12. Once one understands the core intuition, however, the rest is mostly detail.

1.5 THE PLAN OF THE BOOK

This book is divided into four parts. Part I contextualizes the state of polling today. Chapter 2 tells the story of how the survey world has lurched from paradigm to paradigm, typically jolted from a consensus view by failed predictions in presidential elections. Chapters 3 and 4 provide an overview of the

https://doi.org/10.1017/9781108697798.002 Published online by Cambridge University Press

current state of polling. Chapter 3 dives into weighting, an important tool to deal with nonresponse in contemporary polling. Parts of Chapter 3 are technical and can be skimmed by those readers less interested in such details. Chapter 4 broadens the lens to consider the entire range of contemporary polling, including use of nonprobability survey methods. For all the creativity and innovation in modern polling, it is striking the degree to which polling continues to rely on the assumption that nonresponse is ignorable.

Part II of this book builds the case that we need a new paradigm that will encompass modern polling realities. Chapter 5 presents a model that incorporates both ignorable and nonignorable nonresponse without privileging either. At its core is Harvard statistician Xiao-Li Meng's Big Data Paradox: "The larger the data size, the surer we fool ourselves when we fail to account for bias in data collection." That is, if the data collection process is problematic, having a lot of data may not only lead us to the wrong answer, but it may lead us to have misplaced confidence in the wrong answer (Meng 2018).

Chapter 6 fleshes out further implications of this paradigm. One major implication is that *random contact* surveys are incredibly valuable. Random contact surveys are what we mostly have today: People are contacted, and then of the contacted individuals, there is a (presumptively) nonrandom process that governs who actually responds. Random contact surveys do not produce random samples, but the Meng framework makes it clear that they limit the damage that nonignorable nonresponse can inflict.

Part III explains how to fight nonignorable nonresponse. Chapter 7 explains tools that assess the vulnerability of a survey to nonignorable nonresponse. These tools are extremely clever and provide a minimal analysis that anyone can do; they do not, however, diagnose or correct for nonignorable nonresponse and may therefore by skipped by readers eager to diagnose and correct for nonignorable nonresponse.

Chapter 8 introduces the logic underlying models that diagnose and correct nonignorable nonresponse. The key insight is that nonignorable nonresponse can leave traces in survey data, especially if we have the rights kinds of survey data. Understanding this insight helps us appreciate just how distinctive these models are from weighting approaches and how they desperately require certain kinds of data.

Chapters 9 and 10 describe the modern selection model toolkit in more detail. Researchers working not only on surveys, but also on economics and epidemiology have made excellent progress fighting nonignorable nonresponse. Chapter 9 presents a range of parametric and semi-parametric selection models. We see, for example, how one can generate weights that account for nonignorable nonresponse when it exists. Chapter 10 emphasizes how randomized response instruments produce the data that give these selection models the best chance to identify and correct for nonignorable nonresponse if it exists. In order to make sense of the many models and data types, Chapter 11 provides an integrated discussion of the models, their strengths and weaknesses. It culminates with a flowchart that can assist analysts seeking to use these models in practice.

Part IV of this book applies the tools. Chapter 12 describes multiple political polls that were designed to identify and correct for nonignorable nonresponse. The goal is to both explain *how* to implement the tools and reveal *what* the tools can uncover. There are strong signs that polls may provide distorted views of attitudes about race, while also exaggerating partisan divides.

Chapter 13 extends the toolkit beyond political polling to public health, another area where nonignorable nonresponse presents major challenges. It is highly likely that the people who choose to test for a given disease such as Covid will have different probabilities of being sick than the rest of the population, making it hard to generalize from observed testing data. This chapter suggests how randomized response instruments can improve estimates of disease prevalence and how the logic of nonignorable nonresponse can be leveraged to better compare prevalence across jurisdictions.

1.6 WHO SHOULD READ THIS BOOK

Asked to describe the media's understanding of polls, the Cook Political Report's Amy Walter said that "journalists talking about polls are like preteens talking about sex. They know all the words. They talk about it a lot. But they have no idea what they're talking about" (Grynbaum 2019). This description may cut a little close even for us nonreporters; certainly, we all recognize how it is easy to lose the thread of what polls mean when we face a barrage new methods and new results.

This book seeks to help people step back to understand the whole story of how nonresponse makes polling difficult, especially in today's post-random sampling environment. It is particularly useful for people who want to understand the limits of conventional methods and how to overcome them. I'll explain where we are – and how we got here. I'll explain the theory behind polling and what has been lost by ignoring the ironically named nonignorable nonresponse. I'll also show that we are not helpless; it may not be easy in practice, but via simple pictures I will make it clear that nonignorable nonresponse can leave traces and that new approaches offer promising avenues for countering the problem.

If you are a student or a politically engaged citizen, the book will give you deeper knowledge of how polls work, helping you know whether to be convinced or skeptical of the latest poll. If you are a polling professional or an academic researcher, the book will help you address that nagging worry that your weights have assumed away potentially important flaws in your survey samples. I provide specific suggestions for how to improve polls, suggestions that are grounded in theory and a broad academic literature.

The book centers on politics, but is broadly relevant. Nonignorable nonresponse can affect *any* effort to extrapolate from a sample to a population, meaning that researchers in health, business, economics, and demography will recognize the problems and benefit from the new solutions.

The book does not cover all aspects of survey research. As characterized by the total survey error approach, survey research can go wrong in coverage, response, question formulation, insincere or unthoughtful responses, social desirability, and poor analysis (Biemer 2010; Groves 2004). All of these are important topics; the focus here is on one of them, nonresponse bias. I also avoid some interesting, yet fundamentally distinct, topics. For example, political polling is often used to anticipate or explain election outcomes. Doing so requires not only a model translating survey responses to the population at large, but also a model mapping the population at large to turnout (see, e.g., Sturgis et al. 2018). Often the biases in survey response and turnout reinforce each other (because the kinds of people who answer polls are the kind of people who are more likely to vote), but I do not attempt here to model turnout.

This book can be read in different ways depending on your background. I present all the main arguments textually, visually, and mathematically. The purpose of the math is to *simplify* the arguments and reduce clutter, but I appreciate that not everyone thinks in such terms. Rest assured, therefore, that if you do not like equations, you will see a story and a picture for everything that matters.

1.7 CONCLUSION

Polling fulfils a valuable function in society, but has become increasingly difficult in recent years. Modern survey samples are nonrandom samples. This is abundantly true for pollsters who do not randomize contact. It is also true of pollsters who randomize contact but hear from only a few of the people they contact.

Nonrandom samples risk two kinds of bias. Ignorable nonresponse bias occurs when the sample distributions of characteristics differ from known population-level distributions. Nonignorable nonresponse bias occurs when the propensity to respond is related to the item being measured even when we control for demographics. Dealing with ignorable nonresponse is relatively easy. Dealing with nonignorable nonresponse is not. In fact, weighting or generating large samples can aggravate bias when nonresponse is nonignorable.

I argue in this book that we need to build from a starting point that recognizes both kinds of biases. We can then build tools and models that are more robust to the full set of risks facing the nonrandom samples produced by all modern polling. The payoffs can be substantial. Writing *before* the 2020 election, Isakov and Kuriwaki (2020) used Meng's 2018 framework to anticipate how polling would perform. They noted that if nonignorable nonresponse occurred on the order it had in 2016 and if weighting hadn't improved (both true, as it turned out), then the 2020 polls were underestimating Trump by around 1 percent in key battleground states and, perhaps more importantly, were vastly underestimating the uncertainty in the numbers. While even their methods understated Trump's performance, their conclusion written before the 2020 election was prophetic: "Our approach urges caution in interpreting a simple aggregation of polls. . . . Our scenario analysis confirms [Biden's lead] in some states while casting doubts on others."

The point is not that all polls are terrible or that we should never believe polls. For all the problems outlined above, polls often perform well – remarkably well in light of the challenges (Jennings and Wlezian 2018; Silver 2021a). In 2018, Nate Silver famously declared that "the polls are all right," echoing the Who's song "The Kids Are Alright" (Silver 2018). If we think of polls as kids, we should acknowledge that yes, polls are going to school, getting decent grades, and eating their vegetables.

But that doesn't mean the kids are totally fine. Every few elections a packet of heroin falls out of one of their backpacks. We probably shouldn't look the other way. Instead, we need more work, more theory, and more innovation to keep polls from veering down the wrong path.