Taking the Study of Political Behavior Online

Stephen Ansolabehere and Brian F. Schaffner

The Oxford Handbook of Polling and Survey Methods
Edited by Lonna Rae Atkeson and R. Michael Alvarez

Abstract and Keywords

This chapter describes the rise of online surveys as a research tool for social scientists. First it provides an analytical framework for understanding how survey mode matters to social science research. It examines the consequences of the trade-off between quality and cost for an entire research program or literature. For survey methodologists, quality boils down to the ability to test a hypothesis using the survey. Second, the chapter examines the controversy over the use of opt-in Internet polls rather than traditional polls. Recent studies have found that high-quality online surveys produce estimates that can be as reliable as those from traditional polls. Using data from over 300 state-level opt-in Internet subsamples from the CCES, the chapter measures the amount of error in a commonly used approach for conducting opt-in Internet surveys and compares it to traditional probability samples. It concludes by considering how to make wiser choices about survey mode.

Keywords: online surveys, survey mode, survey methodology, survey innovation, survey quality

SURVEY research in the United States has crossed a threshold. Over the past two decades there has been an explosion in the number of academic studies making use of Internet surveys, which are frequently conducted using opt-in samples rather than samples of randomly selected individuals. News media polls have followed suit, and today non-probability Internet polls are nearly as common as random digit dialing phone polls. Internet polling is here to stay, at least until the next revolution in survey research.

The change has been driven by a variety of factors. First, phone surveys have become more difficult to conduct. Since 2005 there has been a precipitous decline in the use of landline phones in the United States, especially among young adults, and there are legal barriers to many techniques used by market researchers for random digit dialing of phone numbers. In addition, social norms about answering phone surveys have changed, causing response rates to most phone polls to drop into the single digits. Second, cost calculations have changed. Survey research firms dedicated to using the Internet and non-probability based sample selection methods as a mode of data collection, such as Knowl-
edge Networks and YouGov, have emerged and have produced relatively low cost alternatives to phone and other modes of survey contact. Third, researchers have realized the survey design opportunities available with Internet polls. Online surveys offer the opportunity to show visuals and videos, to conduct experiments within surveys easily, and to implement new forms of questions. They are also generally easy to field quickly, making them a way in which researchers can receive data back in a timely manner. Fourth, people respond to Internet surveys in several advantageous ways. There is evidence of less social desirability bias when no interviewer is involved, and people read faster than they speak, meaning that people can answer many more questions in an online poll than in one conducted over the phone in the same amount of time.

While some firms like Gfk (formerly Knowledge Networks) deliver surveys online to panels that are recruited with probability sampling methods, most online firms use some form of opt-in recruitment strategy. While techniques often vary widely across online polling firms, the highest quality firms tend to spend substantial resources recruiting individuals to join their panels through online advertising, referrals, and other approaches. Once people join the panel, they are asked to take surveys from time to time, often in exchange for points that can be redeemed for some reward (like a gift card). Some firms, such as YouGov, have millions of individuals throughout the United States who are active members of their panel. When a researcher contracts with YouGov to conduct a survey, the firm attempts to collect responses from a sample of their panelists who would be fairly representative of the population that the researcher is interested in.

While these strategies are often referred to as nonprobability samples, that terminology can be misleadingly simplistic. First, some online polling firms, like YouGov, sample individuals from their panel using an approach that is based on a randomly selected target to which volunteer members of the panel are then matched based on their demographics (see Rivers 2007). Thus, this technique does have grounding in probability sampling. Second, as many scholars have noted, the line between probability and nonprobability recruitment has blurred considerably in the era of exceedingly small response rates. For example, Andrew Gelman and David Rothschild (2014) note, “No survey is truly a probability sample. Lists for sampling people are not perfect, and even more important, non-response rates are huge. . . . Rather than thinking in a binary way of probability vs. non-probability sampling, perhaps it’s better to think of a continuum.”

The point that Rothschild and Gelman are making is that when response rates are less than 10% and others in the population are not included in the sampling frame at all, it becomes much more difficult to treat anything as a pure probability sample. Accordingly, all survey researchers now engage in a substantial amount of modeling (e.g., weighting) to ensure that the sample they ultimately end up with is representative of the population they are attempting to draw inferences about. However, it is typically the case that online opt-in surveys require more modeling than well-designed surveys using probability sampling. We consider this point in greater detail below. However, it is important to keep in mind that surveys do span a continuum in terms of the degree to which they rely on modeling versus random selection. Nevertheless, in this chapter we use the terms online and/
or *opt-in* as shorthand for surveys that rely more on modeling and less on random sampling and face-to-face, telephone, and/or probability samples.

The transition to opt-in, online polls has been controversial in the community of survey researchers (e.g., Voosen 2014). The most obvious opposition comes from incumbent survey organizations: those invested in phone and face-to-face polls. However, as we discuss below, there has also been strong resistance in the scholarly and methodological communities. The shift away from pure random sampling was driven partly by the increasingly higher nonresponse rates to existing survey methods as well as the distinct approach that online surveys required. The new technologies also had to prove their mettle. Could researchers be confident that the new survey methodologies yielded valid estimates of opinions and behaviors? What would be the basis for drawing statistical inferences from samples that were not randomly selected? As the move to online polling occurred—and in the mid-2000s it seemed inevitable because of the opportunities the technology presented and the increasing challenges faced by traditional modes—what would be gained and lost in the transition to online polling?

This chapter examines the trade-offs that the survey research and public opinion field has faced in the transition to online opt-in polling. The heart of the matter is not which mode is right or wrong, good or bad. Rather, the transition that survey research is undergoing forces us to understand how to best make decisions about how research is conducted.

In this respect, the discussion here points to three significant conclusions, which we return to at the end of the chapter. First, transitions take time. The early attempts at Internet polls were error prone, but they improved markedly over time and tend to vary significantly across survey firms (e.g., Kennedy et al. 2016). The field’s understanding of survey method is not, then, static, but evolves with societal, technological, and industry changes. Second, a healthy survey research field will allow for a variety of approaches. The new challenge is not to pick one best approach, but rather how to synthesize information from different approaches. By combining data collected using different approaches we may be able to improve our methods by guarding against the weaknesses in any single approach. Third, there is a need for ongoing testing. We should constantly re-evaluate survey methods, whether they be recently developed or long established. After all, we have learned that the effectiveness of survey methods can wax and wane with changes in technology and society, even if the approach itself remains static.

In the next section we discuss the relationship between quality and cost when conducting survey research. We then turn to focusing on how opt-in Internet surveys stack up both in terms of their overall accuracy and also with regard to the manner in which they are administered to individuals.
Survey Quality and Cost

What has been gained or lost in the transition to online polling? The transition over the past fifteen years from random digit dialing phone polls to opt-in panels that rely on the Internet for response has often been framed as a choice between higher quality probability samples and lower cost (but lower quality) opt-in Internet samples (e.g., Pasek and Krosnick 2010; Chang and Krosnick 2009). That choice was the focus of important literature on mode effects, which we discuss in the following two sections.

The potential trade-off between quality and cost is crucial in research design generally, not just the method through which samples are drawn and surveys conducted. In the scholarship on survey method, researchers have often focused on the total survey error (TSE) approach, which recognizes that various components of a survey combine to affect the total error rate of that survey (e.g., Groves and Lyberg 2010). The resources of researchers—time and money—are limited. With additional resources, it is usually possible to improve on our data collection methods. But given the constraints faced by most researchers, we must decide how to best to allocate our resources. Thus, in this section we consider how to balance the TSE of different approaches with the resources needed to carry out those approaches.

Survey research has transitioned through many different modes, from in-person or face-to-face surveys, to mail surveys, to phone surveys, to Internet surveys, and now, potentially to surveys administered through social media, mobile devices, or services, such as Mechanical Turk. Each transition in survey mode is almost always framed as a choice between high-cost, high-quality methods and low-cost, low-quality methods. In the 1970s, for example, the debate was whether to switch from in-person and mail surveys to random digit dialing phone surveys. At that time, the phone surveys were viewed as suspect, and in-person, face-to-face surveys were taken as sufficiently superior in quality that they must be maintained as the standard methodology for survey research (e.g., Klecka and Tuchfarber 1978; Weeks et al. 1983). But the cost constraints of in-person, face-to-face surveys meant that research organizations could conduct many fewer surveys than they could with phone surveys. In the late 1980s there was an explosion of the use of phone surveys for market and political research because researchers could more quickly field their surveys and could take many more readings of public opinion. In the area of election surveys, for example, the 1988 and 1992 elections saw a rapid increase in the number of election polls conducted by media organizations to gauge the horse race between the Republican and Democratic candidates. The horse-race coverage became a standard part of the story of the election. By the early 1990s, random digit dialing phone surveys had become the new standard.

The control of quality in survey research has traditionally come through the use of random sampling. A 2010 report by the American Association of Public Opinion Researchers (AAPOR) on survey sampling methods stated strongly that random sampling is the indus-
try standard (Baker et al. 2010). That report emphasized concerns about quality, rather than cost, and promoted a specific technical approach to valid survey research.

Why do random sample surveys produce high-quality studies? The Polling 101 version of random sample surveys goes something as follows. A surveyor randomly selects a certain number of individuals from a population: a random sample. By that we mean that all people have a probability of being selected into the sample, and that probability is known and is independent of any characteristic of the individual. That holds true if a device such as a coin toss or a random number generator creates the probability of selection. Further, it is assumed that those selected to participate all respond to the survey and answer questions truthfully and fully.Crudely speaking, that is what is meant by a random sample survey.

The value of this idealized version is that it states a set of assumptions that imply an elegant statistical model of the survey that allows for estimation of and inference about characteristics of a population. More generally, the key assumption underlying the theory of estimation and inference using surveys is that cases are selected into the sample by a process that is independent of any important feature of the sample, also known as the ignorability assumption (Gelman et al. 2004). Randomness in the sample selection process ensures ignorability of the selection (or missingness) of the data, assuming that every individual who is sampled by the surveyor takes the survey.

From the assumption of random sampling, statisticians have developed a theory of estimation and inference. Under the assumption of random sampling (along with complete and truthful response), one can apply the central limit theorem to define the distribution of possible outcomes from a survey and use that distribution to make inferences, such as the degree of confidence in an estimate. So, for example, the typical news story about a poll usually states that a certain proportion of the population has a given characteristic (e.g., approves of the president) and that there is a margin of error of plus or minus 3 percentage points for that estimate. What is meant by that statement is that there is a 95% probability that the true proportion of the population that has that characteristic is within 3 percentage points of the estimate yielded by the survey. Thus, if a poll with a 3 point margin of error finds that 45% approve of the president, then the true value is very likely to be somewhere between 42% and 48% approval.

The random sample survey with complete and truthful response is the proverbial “gold standard” of survey research. Like all proverbs, it has a kernel of truth surrounded by a healthy coating of myth. Perhaps the most troubling problem for conventional random sample surveys has been declining response rates. In other words, a researcher can select a random sample, but the researcher cannot force those sampled to respond. If some types of people are more likely to refuse to participate than other types, then the sample will ultimately be biased. For example, younger adults are often harder to contact and less likely to be willing to respond to surveys, which means that the samples obtained by pollsters are often much older than the population that they are attempting to make inferences about.
The American National Election Study (ANES) expends considerable effort to construct random samples of the U.S. population based on addresses and then to conduct face-to-face interviews. According to the ANES, the response rate to the study has fallen from 80% in 1964 to 60% in 2000 to 53% in 2008 to 38% in 2012.\(^2\) The Pew Center on People and the Press conducts the highest quality phone studies possible. That research organization reports declining, even lower response rates to phone polls. From 1997 to 2012, the response rate to the Pew phone surveys dropped from 36% to just 9%.\(^3\)

The high nonresponse rates associated with phone and face-to-face surveys since the 1990s created substantial doubts about the validity of the survey enterprise, and opened the possibility for another approach. Under the usual statistical theory, high nonresponse rates raise concerns about the confidence in the assumption of pure randomness; after all, most of the people who were randomly selected into the sample have declined to participate. As a result, researchers must either fall back on the assumption of ignorability of nonresponse (i.e., assume that those who refused to answer were no different than those who participated) and noncoverage (i.e., people who cannot be reached through the survey mode) or attempt to adjust the survey at the data analysis stage to correct for patterns of nonsampling errors that are nonignorable (i.e., by weighting the sample). That is, researchers either had to believe that the 60% of people who refused to respond to the ANES in 2012 were no different than the 40% of people who did respond, or they had to use statistical methods to “fix” the sample to make those who responded look like the original random sample.

Even before the transition to online polling began, survey researchers were already using weighting to deal with the challenges faced by plummeting response rates. This is not to say that the actual quality of surveys had declined. Rather, the key facts about declining response rates had led to an increased impetus among survey researchers to use statistical methods to adjust for the fact that samples violated the ignorability assumption. These rising concerns about sampling also provided an opening for survey innovation, a search for alternative modes and new ways of thinking about survey design. The challenge for new modes, such as the opt-in Internet survey, was demonstrating that these new approaches were of sufficiently high quality and lower cost to justify the move. The main concerns were nonresponse, noncoverage, and the lack of randomness as a protection against idiosyncratic errors in sample selection.

The costs of surveys can vary considerably across modes and even within modes. A typical Internet sample of 1,000 respondents costs in the neighborhood of $10 to $20 per interview. Special samples (say of a specific demographic or region) can be considerably more expensive.\(^4\) The costs of a random digit dial phone poll are typically at least 50–100% higher than high-quality Internet polls of the same population.

The most expensive surveys, by far, are address based samples conducted face-to-face, such as the ANES and the Panel Study of Income Dynamics. The ANES reports that the cost of fielding its survey (excluding other activities associated with the project) was approximately $3 million for 2,000, or a staggering $1,500 per interview. The possible costs
of a national survey of American adults, then, can range from approximately $10 per interview to more than $1,000 per interview.

How should we think about the trade-off between cost and quality? What are the benefits of a high-quality survey, and what are the losses associated with a lower quality survey? Quantifying those benefits and losses is essential in making a systematic choice about research design.

Typically, the trade-off between quality and cost is considered only in relation to a single study. Given a fixed amount of money, a research team chooses a survey mode and designs and implements its questionnaire. And in preparing a grant, a research team must justify its choice of survey methods and modes. In making design decisions, researchers must consider the consequences of making either Type I or Type II errors. That is, they must weigh concerns about their wrongly concluding that a hypothesis is correct when in fact it is not, or wrongly concluding that a hypothesis is wrong when in fact it is true.

While researchers typically make decisions about mode in relation to a single study, in academic research it is more fruitful to think about the quality-cost trade-off not in terms of a single survey but in terms of a series of studies that all seek to answer the same question—that is, in terms of an entire literature. If a discipline chooses a higher quality methodology, then scholars can answer a given question or test a given hypothesis or conjecture more efficiently than if the discipline used less accurate methods.

Suppose we conduct one study under the strong assumptions of random sampling, with 100% response rate and no misreporting. We use this survey to produce a point estimate (say approval for the president) and a confidence interval. In that case, the chances of “getting the answer right” (creating a confidence interval that includes the true population value) are 95% for a traditional level of confidence. We take that as a baseline.

One way to quantify the loss associated with an inferior methodology is to ask how many studies researchers would have to do to reach the same conclusion as a high-quality survey with 95% confidence. There are many ways to quantify that specific criterion. Suppose that we use simple majority rule: Do a majority of studies confirm or disprove a given estimate or hypothesis? Adding a degree of confidence to that statement, we seek to establish how many studies of inferior quality researchers would have to conduct to have a 95% probability that a majority of studies reach the correct conclusion. We think of this as a quantification of what is meant by a consensus in a scientific community.

Take, as an example, two types of studies. One type of study uses the superior methodology (random sampling, complete and correct responses). From this, one can build a confidence interval or conduct a hypothesis test that, in a classical statistical framework, will have a .95 probability of being true. This is our baseline criterion. The other type of study uses an inferior methodology. Suppose that the inferior approach would confirm a hypothesis, if the hypothesis is true, with probability .9 (rather than .95).⁵
How many studies of inferior quality must be done to have 95% confidence that the body of research arrives at the right result? Assume that a series of three independent studies is conducted using the inferior methodology. The probability that all three studies confirm the hypothesis is .729 (.9 × .9 × .9), and the probability that two of the three confirm the hypothesis is .243. Thus, the probability that a majority (two or three) of the three studies confirm the hypothesis correctly is .972. Importantly, this calculation assumes that the studies are independent of one another. Positive correlation among the studies can make this an underestimate of the number of studies needed; by the same token, negative correlations among studies can actually gain efficiency. Setting that concern aside, under the assumption of independence, if we conduct three inferior studies, we have as much confidence that a majority of those studies are correct as we would if we conducted one study using the superior methodology.

This approach allows us to quantify the quality-cost trade-off. A direct cost calculation is simply the number of surveys that are required to obtain a consensus, given a level of quality of a survey. Other considerations, such as opportunity costs of researchers, might be factored in as well. The simple implication of the calculation above is that it is worth using the superior quality survey only if the cost of doing one such survey is less than the cost of doing three inferior quality surveys. Likewise, it may be worth using an inferior survey methodology if the cost of such surveys is less than one-third the cost of the superior methodology. We see a similar logic play out when it comes to horse-race polling during campaigns. While some very high-quality surveys are useful indicators of the state of the race in their own right, most seasoned scholars and pundits focus on aggregated indicators of the state of the race taken from multiple polls (i.e., polling averages). The justification for this approach is that one can generally learn at least as much from averaging multiple inferior polls as from looking at a single poll, even one of very high quality.

Viewed in this way, it becomes extremely useful to measure the relative quality of various survey methodologies to contextualize the cost differentials. Denote the degree of quality of the inferior methodology as q, the probability that the hypothesis is confirmed using the inferior quality methodology given that the hypothesis is right. In the calculation above, q = .9, and we ask how many studies with q = .9 must be performed to have a probability of .95 or higher that the majority of those studies confirm the hypothesis when that hypothesis is true. Now consider doing the same thought experiment for lower levels of quality, namely, q = .8, q = .7, and q = .6.

Table 4.1 presents the number of studies needed in a literature to attain a 95% level of confidence that a majority of the studies conclude that the hypothesis is true, when in fact it is true. Again, we assume that the studies are independent of one another. Lessening the level of quality from q = .9 to q = .8 increases the number of studies needed to reach at least a 95% level of confidence from three to seven. In other words, if survey quality concerns raise the probability of a false negative from .05 to .20, then a research community must conduct at least seven studies before a sufficiently strong consensus is reached. Continuing in that vein, if the level of quality drops to q = .7, then the research community must conduct at least fifteen studies to reach a consensus, and if the level of
quality is as low as \( q = .6 \) (meaning there’s a 40% chance of a false negative on any single survey), then the research community would have to conduct sixty-five studies before a majority of studies clearly answers the research question.

This formalization of the quality-cost trade-off has several important implications for the choice of survey mode.

<table>
<thead>
<tr>
<th>Probability Correct*</th>
<th>Number of Studies Needed</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q = .9 )</td>
<td>3 (at least 2 of 3 correct with probability .95)</td>
</tr>
<tr>
<td>( q = .8 )</td>
<td>7 (at least 4 of 7 correct with probability .95)</td>
</tr>
<tr>
<td>( q = .7 )</td>
<td>15 (at least 8 of 15 correct with probability .95)</td>
</tr>
<tr>
<td>( q = .6 )</td>
<td>65 (at least 33 of 65 correct with probability .95)</td>
</tr>
</tbody>
</table>

(*) Probability that one will conclude \( H \) is true, given that it is true.

First, very high-quality surveys have a significant edge in choice of research method. A small degradation of quality, say from \( q = .95 \) to \( q = .90 \), assuming independence of surveys, means that multiple studies must be conducted to test with high confidence a hypothesis, or that sample sizes must be increased considerably. In other words, a survey that has a 10% error rate imposes three times as much cost (three times as many studies need to be done) as a survey that has a 5% error rate. The cost, in terms of the total number of studies required to achieve a consensus, grows exponentially as the rate of false negatives grows.

Second, the lower cost of Internet polls has been winning out over the “gold standard” polls, in part because of the exceptionally high cost of address based sampling, face-to-face polls. Consider the comparison of the costs of conducting the ANES in 2012 and a similar-sized Internet poll. The ANES’s face-to-face, in-person survey is more than one hundred times more expensive to do than a high-quality Internet poll. In other words, for the cost of the ANES one could do at least one hundred high-quality Internet polls. With that cost differential, it is worth it to the scientific community to use the lower cost modes.
to answer research questions, even when the probability of a false negative is as high as 40%!

Third, this framing of the problem raises the natural question of what constitutes a scientific consensus. Is a .95 probability that a test confirms the hypothesis when that hypothesis is true too high? Might a research community feel that a consensus emerges with a lower probability that a majority of studies reach the same conclusion? If a consensus emerges at a lower level of confidence, then the advantage of the higher quality approach is even less pronounced.

The approach we have sketched here also offers insight into the question of multiple or mixed modes of survey research. Suppose a research group conducts three surveys to test a hypothesis. That research group might conduct three superior quality surveys (at considerable expense) or three inferior quality surveys (at much less cost), or it might employ a mix of approaches. An analogous calculation to that in Table 4.1 reveals that there may be an advantage to mixing the modes, or, equivalently, using multiple survey modes in a research project or literature.
### Table 4.2 Survey Quality, Mixed Modes, and the Probability That a Majority of Studies Reach the Correct Result

<table>
<thead>
<tr>
<th>Quality of Inferior Survey</th>
<th>3 Superior Quality Surveys</th>
<th>2 Superior 1 Inferior Quality</th>
<th>1 Superior 2 Inferior Quality</th>
<th>3 Inferior Quality Surveys</th>
</tr>
</thead>
<tbody>
<tr>
<td>q = .9</td>
<td>.993</td>
<td>.988</td>
<td>.981</td>
<td>.972</td>
</tr>
<tr>
<td>q = .8</td>
<td>.993</td>
<td>.979</td>
<td>.944</td>
<td>.896</td>
</tr>
<tr>
<td>q = .7</td>
<td>.993</td>
<td>.969</td>
<td>.889</td>
<td>.784</td>
</tr>
<tr>
<td>q = .6</td>
<td>.993</td>
<td>.960</td>
<td>.816</td>
<td>.648</td>
</tr>
</tbody>
</table>
Table 4.2 presents the probabilities that a majority of studies reach the correct conclusion, for various levels of survey quality and mixes of inferior and superior methodologies. If the researchers were to conduct three superior quality surveys, each of which has a .95 probability of concluding that the hypothesis is correct when in fact it is, then there is a .993 probability that at least two of three or three of three surveys reach the correct conclusion. Interestingly, if the researchers were to include an inferior quality survey along with two superior quality surveys, they would have nearly the same (very high) level of confidence that a majority of their surveys are correct. If \( q = .9 \) for one of the surveys and \( .95 \) for two of the surveys, then the probability of a correct conclusion among a majority of surveys is .988. Even if the low-quality survey has a \( q \) of just .60, the probability that a majority of the three surveys is correct is .960. See the third column of the table. Using multiple surveys provides some protection against false inferences from inferior quality surveys. That said, quality does harm inference: the lower the quality, the lower the probability of reaching the correct inference. The drop-off in confidence can be quite large with lower quality, especially when all surveys are of the inferior sort.

One important implication of the simple analysis is that not all surveys need to have the same quality for a scientific consensus to emerge. For example, with one superior quality survey and two inferior quality surveys (\( q = .8 \)), the probability that a majority of surveys yields the correct answer is still approximately .95.

This points to a possible cost-saving approach in research. Having multiple survey modes allows a research group or an entire field of study to lower the cost of reaching a scientific consensus. In fact, having all surveys be of very high quality might even be inefficient. If, to reach consensus, at least three studies need to be conducted in a literature, then three very high-quality surveys will have an extremely high probability of agreement. A discipline, then, can tolerate a mix of higher quality and lower quality surveys and still attain a high probability that a majority of surveys reach the correct conclusion.

Having multiple survey modes in a research literature also allows for testing of the validity and quality of the modes. If there are several modes being actively used, researchers can compare the relative quality of various modes.

As new modes for conducting surveys emerge, the key question, then, is what quality of results is derived from those new modes. How closely do new modes of conducting surveys approximate the ideal of a random sample survey with complete and truthful answers? In the early 2000s, as nascent Internet survey methods began to emerge, that is precisely the question survey researchers faced. And today, as Mechanical Turk and other platforms for data collection emerge, the same questions arise.

Quantifying the Quality of Internet Surveys

There are two approaches to quantifying the quality of a result estimated from a particular method: (1) compare the survey results from that mode with objective indicators and (2) compare estimated quantities (means, variances, correlations, and regression...
coefficients) for identical questions asked in different survey modes (e.g., phone versus mail or phone versus Internet).

Comparison with objective indicators offers the strongest measure of survey quality because it allows researchers to compare their survey estimates with the quantity that they are actually trying to estimate. Suppose that a survey attempts to measure a characteristic of a population, such as the percent of votes won by the Republican candidate for president in each of several states. The survey samples a few hundred people in each state and asks for whom they voted for president. The deviation between the survey estimates and the actual election results (the true or population value) reflects the random and non-random errors that occur in the survey process. This is often referred to as the TSE (Platek and Sarndal 2001). Total survey error includes the deviation of the estimated value from the actual population value as a result of all parts of the survey, including non-response, misreporting, poorly asked questions, and other problems. These errors may be random (and add to the variance of the estimate) or systematic (and cause bias in the estimates). In statistical terms, the TSE is the mean squared error, which equals the square of the bias of the survey estimate of a given quantity (e.g., a mean or proportion or regression coefficient) survey plus the sampling variance of the estimated quantity (i.e., the square of the standard error).

To measure TSE, or mean squared error, multiple measures are needed. The deviation of any one survey’s estimate from the actual value of a given quantity is a single realization of the TSE. Suppose the same survey method is repeated many times (either many different quantities within a single survey or many replications of the same survey), and the deviation of the survey from the actual value is calculated for each replication. The average of those deviations gauges the bias—the extent to which the survey instrument is systematically too high or too low—and the variance of the deviations estimates the mean squared error.

The Cooperative Congressional Election Study (CCES) provides an ideal example and case for measuring the TSE associated with Internet polls. The CCES is conducted every year and is designed to measure the vote choices and political preferences of American adults. The study employs very large samples, in excess of 30,000 in 2006 and 2008 and in excess of 50,000 in 2010, 2012, and 2014. The large samples make it possible to estimate the vote in each state for president, U.S. Senate, and governor, and to compare those estimates to the actual results at the state level. For each state one can calculate the theoretical (or expected) standard error under the usual assumptions of sampling theory, such as random sampling or ignorability, and one can calculate the deviation of the estimate from the actual result. The average deviation (bias), mean squared error, average number of cases per state, and expected standard error are presented in Table 4.3 for each statewide race and year for which the CCES provides estimates (see Ansolabehere and Schaffner 2015, 16–20). In all there are twelve separate contests, but each is measured at the state level. Consequently, there are over three hundred individual-level elections (for each unique combination of state, year, and office) represented in Table 4.3. The
Table displays the results aggregated to each year and office and, at the foot of the table, aggregated across all states, years, and offices.

Table 4.3 presents an overall picture of the accuracy or quality of the CCES, relative to the ideal survey of the same size. The average bias is 0.4%, which means that averaging over every case, the average deviation overstated the Democrat’s share of the vote, but only by four-tenths of 1 percentage point. The average mean squared error is 3.19%, and we contrast that with the expected standard error. The average standard error, under the assumption of ignorability or random sampling, is 2.36%. That is approximately 25% smaller than the mean squared error, our estimate of the true variance of the TSE.
## Table 4.3 Comparing Survey and Actual Results: Bias, Mean Squared Error, and Standard Error for the Cooperative Congressional Election Study, 2006–2014

<table>
<thead>
<tr>
<th>Year</th>
<th>Position</th>
<th>Average Error (Dem. Bias)</th>
<th>Root Mean Squared Error</th>
<th>Average Number</th>
<th>Expected Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Governor</td>
<td>−0.84%</td>
<td>3.95%</td>
<td>626</td>
<td>3.59%</td>
</tr>
<tr>
<td></td>
<td>U.S. Senate</td>
<td>+0.34%</td>
<td>3.38%</td>
<td>515</td>
<td>4.26%</td>
</tr>
<tr>
<td>2012</td>
<td>President</td>
<td>+2.16%</td>
<td>3.32%</td>
<td>1,069</td>
<td>1.53%</td>
</tr>
<tr>
<td></td>
<td>U.S. Senate</td>
<td>+2.27%</td>
<td>3.66%</td>
<td>1,217</td>
<td>1.94%</td>
</tr>
<tr>
<td></td>
<td>Governor</td>
<td>+1.49%</td>
<td>3.56%</td>
<td>666</td>
<td>1.43%</td>
</tr>
<tr>
<td>2010</td>
<td>Governor</td>
<td>−0.95%</td>
<td>2.22%</td>
<td>982</td>
<td>1.93%</td>
</tr>
<tr>
<td></td>
<td>U.S. Senate</td>
<td>−0.50%</td>
<td>1.30%</td>
<td>882</td>
<td>1.98%</td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Taking the Study of Political Behavior Online

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>President</td>
<td>+0.57%</td>
<td></td>
<td>940</td>
<td></td>
<td>2.05%</td>
</tr>
<tr>
<td>U.S. Senate</td>
<td>−0.58%</td>
<td></td>
<td>638</td>
<td></td>
<td>2.26%</td>
</tr>
<tr>
<td>Governor</td>
<td>+1.12%</td>
<td></td>
<td>511</td>
<td></td>
<td>2.35%</td>
</tr>
</tbody>
</table>

### 2006

<table>
<thead>
<tr>
<th>Position</th>
<th>Change 2006</th>
<th>2006 Votes</th>
<th>2006 Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Governor</td>
<td>−0.04%</td>
<td>604</td>
<td>2.53%</td>
</tr>
<tr>
<td>U.S. Senate</td>
<td>+0.28%</td>
<td>689</td>
<td>2.42%</td>
</tr>
<tr>
<td>Average</td>
<td>0.43%</td>
<td>2.36%</td>
<td></td>
</tr>
</tbody>
</table>


A further analysis of the data allows us to calculate the quality of the CCES along the lines of the analysis suggested by Table 4.1. For each office, year, and state we calculate the squared deviation of the survey result relative to the squared standard error for that state’s sample. The average of those relative deviations estimates the expected quality of the survey. It is a multiplier indicating how much larger the true variance of the survey is (the variance of the TSE) than the variance of the idealized survey. That calculation suggests that the true standard deviation of the survey is approximately 1.35 times the expected (p. 88) standard error. We can now use that standard error to construct a test statistic, rather than the conventional standard error calculation. The implication is that this Internet survey lowers the quality of inferences somewhat. If the probability of a false negative is .05 for a test statistic constructed using the usual (expected) standard error, then the probability of a false negative is .15% using a test statistic constructed using the estimated square root of the mean squared error as the standard error. A more appropriate calculation of the quality of the survey relative to the ideal standard is of the estimated mean squared error relative to the expected standard error for each office and year.

Is that a substantial degradation compared to other surveys? Very few other surveys calculate the TSE associated with their projects. An analogous, but not as expansive, concept is the design effect, which measures the variation in a survey that comes from clustering and nonindependence of observations and other features of the design that can produce higher sampling variances than occur with pure random sampling. The design effect does not capture biases that occur due to misreporting, nor does it account for whatever bias remains after weights are applied to adjust for nonresponse. The design effect, however, can be thought of as degradation in quality relative to pure random sampling, as such effects increase the probability of false negative inferences. The design effect of the ANES has been estimated to be in the neighborhood of 1.2 to 1.6. In other words, the inflation of the standard error with the Internet sample used by the CCES is approximately on the same order as the design effect associated with the sampling procedure used by the ANES. This suggests that there may be little degradation in the ability to draw inferences using Internet polls relative to traditional random sample, face-to-face surveys.

These calculations are presented to demonstrate how researchers may assess the quality of new survey methods relative to existing methods. In the case of the YouGov samples relied on by the CCES, there is little evidence of systematic bias and evidence of some loss of precision relative to the idealized pure random sample survey. However, no surveys currently match the idealized pure random sample survey. Comparing the ANES design effects and the TSE of the CCES, there appears to be little degradation in the ability to draw inferences compared with more traditional sampling modes.

Any new and untested methodology faces tough scrutiny, and ought to. Total survey error provides a general framework for assessing the quality of new modes of surveying. This framework allows us to measure in clear quantitative terms the quality side of the cost-quality trade-off. The example of the CCES offers a good case study of the use of TSE to provide a critical evaluation of the performance of an Internet survey. Importantly, the analysis of the CCES over a ten-year time frame revealed that the study did not, in fact,
represent a significant reduction in quality, compared with the design effects of tradition­al surveys.

A second manner of assessing the quality of any new approach relative to established methodologies is a carefully designed study that compares the modes. Unlike TSE, the framework of a mode study is to compare the estimates yielded by competing modes. No comparison with an objective reality is usually made, so it is possible that there are biases that affect all modes. However, a mode study is useful in determining whether a new mode might alter conclusions we have drawn using established modes.

Studies of mode differences in the early 2000s found substantial differences between opt-in Internet samples and random digit dialing phone samples. For example, in a study of alcohol use among young adults, Link and Mokdad (2005) found that phone and mail surveys yielded similar results, but that their Internet sample produced different results. Studies such as this one led a group at AAPOR to conclude in 2010 (Barker et al. 2010) that opt-in Internet samples differed from other modes of inquiry.

More recent research, however, shows few or no significant differences between traditional modes and opt-in online survey approaches. Ansolabehere and Schaffner (2014) conducted a mode study comparing phone, mail, and Internet samples. They found no substantial differences across modes in reported behaviors, such as voting, vote preference, donating blood, smoking cigarettes, moving, or owning a home. They found no significant differences in regression coefficients or correlations across modes in explaining approval of Obama; approval of Congress; and attitudes about abortion, affirmative action, gay marriage, Social Security privatization, or taxes. We have also conducted mode studies comparing the face-to-face version of the ANES with versions conducted in two separate online survey formats. Our results show that in terms of both point estimates and cross-item correlations, online surveys track closely with responses secured through the face-to-face sample.

Other studies have reached similar conclusions. The Pew Center for People and the Press conducted a study in 2015 comparing a survey recruited through random digit dialing with nine opt-in Internet samples (Kennedy et al. 2016). That study found that the randomly selected sample was exactly in the middle on most of the measures gauged. The authors concluded that vendor choice matters much more than mode. Other recent studies reach similar conclusions: the differences between quality opt-in and random digit dialing samples have become trivial. What is most important is not which mode you use, but choosing a high-quality vendor to execute the selected approach.

Whether the standard is the absolute level of quality (TSE) or the relative level of quality, the past fifteen years have witnessed a substantial improvement in the demonstrated quality of opt-in Internet surveys. Over this time span Internet surveys turned a significant corner. Although some concerns remain about the use of opt-in surveys, high-quality Internet surveys appear to have gained broad acceptance both within and beyond academia.
Taking the Study of Political Behavior Online

One important lesson from the debate over the quality of Internet surveys is that quality control is essential for surveys, which are a vital research tool. The need to maintain quality creates a difficult problem for those organizations that attempt to set standards for the field, such as AAPOR. Standards seem necessary to maintain quality, but they also can stifle innovation and the evolution of the field.

Qualitative Differences in the Changing Modality of Surveys

An entirely separate set of issues drove the rise of Internet polls over the past decade: the qualitative differences between an online interface and interviewer-led questions. Online polls present new opportunities for conducting research, including the ability to show respondents videos and images and to present new question formats. Experimentation has driven the move online as much as considerations of cost and quality (Evans and Mathur 2005; Sue and Ritter 2012). While the quality-cost debate concerns the validity of population estimates compared with random sample surveys, the opportunity to conduct experiments has made the Internet survey a natural choice. For at least a generation, psychologists (working in all fields) have relied on experiments involving college students to test their ideas. Internet surveys (and newer tools such as Mechanical Turk) offer a much broader population with which to conduct research.

In this respect we see three different attitudes regarding modality. First, Internet surveys differ in mode of data collection only. The quality-cost trade-off treats the questionnaires as the same and merely views the Internet as a more convenient and less expensive mode of data collection. This is an important aspect of the choice of mode, as researchers do not want to lose the continuity with past research, especially for long-lived research projects like the ANES, the General Social Survey, or the Panel Study of Income Dynamics.

A second view, to use Marshall MacLuen’s phrasing, is that the medium is the message. The rise of the Internet and social media has fundamentally changed the way people communicate. The rise of online polling is simply the adaptation of research on social attitudes, opinions, and behaviors to changes in technology and society. The random digit dial phone survey itself was an adaptation to changing communications in society. Sticking with the landline-based mentality today amounts to sticking with older ways of communicating, which are quickly becoming inadequate for the study of society. By 2012 one-quarter of all people could not be reached by a random digit dial phone survey. That number is estimated to exceed one-third of all people in 2016, and it will continue to grow. Many more people are now accessible online.

Not only do new media reach more people, but they involve fundamentally different forms of communication. The Internet is a visual medium. Respondents read online surveys, rather than have the surveys read to them by an interviewer, and the removal of the interviewer from the process makes progressing through the survey much quicker. Visuals and
video can be embedded in a survey, and it is easier to randomly assign survey respondents to see different versions of a message or image. These innovations have opened new ways of asking questions and new ways of analyzing data. The length of time it takes to answer a question, for example, can be easily recorded and provides implicit measures of the degree of cognitive effort a respondent expends in answering a question (Mulligan et al. 2003).

People also interact with new media differently, and that too is part of the survey experience. For example, the 2008 ANES face-to-face survey included a seventy-three-minute pre-election interview and a ninety-one-minute post-election interview. These were some of the longest conversations that respondents had about politics, especially with a stranger. Most Internet surveys also allow respondents the flexibility to complete questionnaires at their own convenience and at their own pace. This means the survey is much less of an intrusion on an individual’s daily routine. Indeed, we have found that people who answer online polls frequently do other things while they are working through the survey (Ansolabehere and Schaffner 2015). For example, 15–20% of respondents watch television while they answer an online survey. Many respondents also take breaks to have a conversation with a family member or roommate, to check email, or to have a phone call. About half of online respondents report doing at least one other thing during the course of taking their survey. Notably, the interruptions and multitasking do not appear to degrade the quality of responses given by respondents.

The self-administered nature of online surveys not only provides a benefit by allowing respondents to finish them at their own pace, but it also means that the responses given are likely to be more accurate. Studies consistently find that respondents are more honest when they answer self-administered surveys, especially those conducted online (Chang and Krosnick 2009; Kreuter, Presser, and Tourangeau 2008). The presence of an interviewer (either in person or on the phone) often discourages respondents from answering sensitive questions truthfully, but when those same individuals can complete the questionnaire privately, they are more likely to provide honest responses.

Overall, online surveys provide an innovative and flexible interface for collecting data. Thus, Internet polls can collect a wider array of data more efficiently, more conveniently, and more accurately than modes that necessitate the presence of an interviewer. When combined with the increasing accuracy and continued affordability of Internet surveys, the flexible and convenient interface is yet another reason that scholars have increasingly used online polls.

Making Wiser Choices about Survey Mode

Survey modes and methods will continue to change as communications technologies change. Today, online polls have gained wide acceptance, and the ascendancy of this new, less expensive methodology has put enormous pressure on more expensive modes, especially face-to-face surveys. And so it goes. New, cheaper ways of conducting surveys replace the old approaches, only to eventually be replaced themselves. Survey researchers
Taking the Study of Political Behavior Online

are trying to figure out the best way to gauge public opinion using mobile devices and social media. Amazon.com’s Mechanical Turk is quickly emerging as a faster and less expensive platform for conducting experiments that were previously done in conventional surveys. And as with other new methods, the debate over the quality of that approach has already begun (e.g., Berinsky, Huber, and Lenz 2012). Mechanical Turk and other new ways of studying political and social behavior will become accepted, and possibly even ascendant. Researchers, facing the inevitable question of how to most efficiently conduct their inquiries, will eventually abandon older methods in favor of newer ones. That cycle of innovation is inevitable. It is a cycle of creativity: new technologies introduce new ways of reaching people, asking questions, and studying behavior.

(p. 92) We have sought in this chapter to introduce a different way of thinking about the future, about what comes next. The debate over methods of studying behavior is often framed in two ways. Both are informative, but neither is adequate. First, the debate is often over “this approach or that”; phone or Internet, mail or face-to-face, probability or opt-in. While that may be the choice that any researcher faces in designing a specific study, it does not reflect the broader concern of a research literature. That quest is to find the most efficient way for research to come to a scientific consensus over an important conjecture or problem. No single mode may be the answer.

Second, the debate over mode is often framed as a debate over a scientific or industrial standard. What are the technical specifications that all researchers must adhere to in order to gain acceptance by a broader community? Such technical specification standards are really social norms, as much as actual quality guarantees. In that regard, it is very important to note that researchers in the United States and the United Kingdom adhere to very different technical standards for their surveys. The technical specification standard that was the norm in the United States for several generations was random sampling; that is, the method of selection must be unrelated to any characteristic of individuals. The technical specification standard that has long been the norm in the United Kingdom is representative sampling; that is, the sample ought to represent the population along several key characteristics, such as age, gender, and education level. If a random sample in the United Kingdom is not sufficiently representative, it is unacceptable. If a representative sample is presented at the AAPOR, it is suspect because the sample was not randomly drawn. These are norms that imply a way that surveys must be done to be acceptable. Such standards serve as a barrier to innovation and barriers to entry in the marketplace of survey research and marketing firms.

Our framing of the problem is that quality goals, rather than technical specifications, are essential. From a scientific perspective, the choice of survey research mode weighs two considerations, cost and quality. If researchers can get higher quality at the same cost, they should buy the higher quality mode. That approach is good not only for the individual researcher working on a tight budget, but also for the scientific community as a whole, as that approach will lead more quickly and efficiently to a consensus around the correct conclusion.
However, we do not operate in a world in which the highest quality is the cheapest. We usually have to make a choice between an expensive but accepted “gold standard” and a cheap but innovative methodology or technology. First, there is considerable uncertainty about the quality of new technologies. Established technologies and technical standards are in place because they won the last fight over methodology. And the research needed to assess the quality of competing modes has rarely been conducted when a new methodology is just emerging, when the choice between competing modes is most difficult to make. Second, the economics of the survey business (or any business) often create a cost difference. The incumbent methodologies are often most expensive because newer technologies are adapted as innovations in cost and quality and because technical standards protect incumbent firms (creating a monopoly advantage for those firms).

If the real goal is maximizing the efficiency of the scientific enterprise rather than conforming to technical standards, how should researchers think about the choice of which methodologies to use now? The framework we have introduced offers guidance about a more effective way of proceeding, both the way to develop a healthy research enterprise and some cautions.

First, there is a place for standards. Standards can offer a means of quality control for the entire research program or research activity. As our examination of Table 4.1 revealed, a few very high-quality studies can be worth dozens of low-quality surveys. The high-quality studies would, then, be the best research method if the costs were not exceedingly high relative to the lower quality methods.

Second, to assess quality there needs to be continual study of new research modes. Researchers cannot make informed design decisions unless data are available about the quality of inferences made using different methodologies. Some of that information can be gained as studies are conducted. For example, the CCES builds in measures of quantities that allow for calculation of TSE. Some of that information can be gained by conducting carefully designed mode comparison studies.

Third, a mix of modes offers efficiency advantages. There is naturally a mix of modes in a research area. At any time there are new ways of conducting research, and those new ideas are contending with established approaches. Our examination of Table 4.2 revealed that a mix of modes can be very efficient, allowing an entire field of research to reach a scientific consensus at a much lower cost. Further, having several different methodologies at work in a field of study allows researchers to compare the different approaches and to draw their own conclusions about quality and the trade-off between quality and cost. Also, different modes can have different strengths and different weaknesses. Using many different modes in a research literature can offer a hedge against the weaknesses of any single mode. Not every survey has to employ a mixed mode, but a healthy literature has a mix of modes across studies. We should be suspicious of anyone who avers that there is one and only one ideal way of doing research.
Fourth, technical specifications of the “gold standard” survey, although they serve an important function, can be a poor way of ensuring an efficient development of scientific understanding. Technical specifications can force the trade-off between quality and cost to be made in one way for all researchers. If every survey is forced, by virtue of technical specifications, to have the same mode, then the advantages of deploying multiple modes are lost.

Fifth, survey quality can become a public good. Research occurs in a decentralized fashion. Every research project makes its own decisions about how to conduct surveys, how to trade off cost against quality. Technical standards can force all researchers to make the trade-off in the same way, say toward the high-quality, high-cost method, but in a way that stifles innovation. The opposite problem can emerge as well. If a research area employs multiple modes of study, there may be a race to the bottom. Every research team might choose the low-cost approach and let someone else bear the cost of the very high-quality study. As a result, a field might collectively proceed very slowly and inefficiently if every researcher chooses the cheap, low-quality option.

The challenge, then, is how to push the boundary outward, how to create innovation in survey quality and survey cost simultaneously. In some respects that already happens. Internet surveys opened new ways of measuring opinions, attitudes, and behaviors. Professional association standards can be helpful in creating guidance about where quality improvements are needed and possible with existing technologies and in maintaining a mix of methodologies so that the rush to a new methodology does not completely lose the value of what existed before. Government agencies, such as the National Science Foundation and the National Institutes of Health, and private research foundations, such as Pew, can serve an important purpose as well. They simultaneously maintain those projects and methods deemed to be very high quality by a scientific community and invest in new technologies and methodologies that show promise of emerging as a platform for the research community at large, such as Time-share Experiments in Social Sciences. And in this respect, there is also tremendous value in careful research about survey methods and robust academic debate about those methods. Professional standards, government and foundation investment, and academic research about survey mode, however, should not be about picking winners, but about informing researchers generally about the quality, the strengths and weaknesses, of alternative ways of studying public opinion and political behavior.

References


Taking the Study of Political Behavior Online


Pasek, J. and Krosnick, J. A., 2010. Measuring intent to participate and participation in the 2010 census and their correlates and trends: comparisons of RDD telephone and non-


Notes:


(4.) Personal communication with Samantha Luks, Senior Vice President, YouGov, San Francisco, CA.

(5.) There are other ways to formalize this choice. For example, it is analogous to the question of how many studies we need to include in a meta-analysis (Valentine et al. 2010). Here we focus on this simple approach, as it makes clear the quality-cost trade-off.

(6.) The CCES is a large-N cooperative survey project carried out every fall since 2006. The survey is conducted by YouGov, using its methodology of matching opt-in respondents...
Taking the Study of Political Behavior Online
to a randomly selected target sample. More details about the survey and access to the survey data can be found at http://projects.iq.harvard.edu/cces/data.


(8.) This comparison actually favors ANES, as the design effect captures the inefficiency in the standard errors to reflect clustering and other features of design, while total survey error contains both inefficiency and bias.


Stephen Ansolabehere
Stephen Ansolabehere is the Frank G. Thompson Professor of Government at Harvard University where he studies elections, democracy, and the mass media. He is a Principal Investigator of the Cooperative Congressional Election Study, and his principal areas are electoral politics, representation, and public opinion.

Brian F. Schaffner
Brian F. Schaffner is the Newhouse Professor of Civic Studies at Tufts University. His research focuses on public opinion, campaigns and elections, political parties, and legislative politics.