45

Probability Versus Non-Probability Methods

Gary Langer

This chapter presents a high-level discussion of the importance and uses of probability sampling in comparison with alternative methodologies. Let's start with a disclaimer: I have no financial interest in any particular sampling methodology. My interest simply is to provide my clients with data that we can obtain as quickly and cost-effectively as their needs demand, but also only with data in which they, and I, can be highly confident. New methods are available. How much confidence we can place in them is the question we face.

Our focus, then, is on survey practices, empirical evaluation, and, ultimately, data quality. An ancient parable speaks to the fundamental importance of this discussion. It tells about about a man who goes to his rabbi seeking forgiveness for having spread false rumors about an acquaintance. The rabbi wordlessly leads his visitor outside to the yard. There he proceeds to rip up a down pillow, then stands silently for a moment until a wind kicks up out of nowhere and sends these tiny feathers swirling in all directions – across the yard, over the fence, through the trees and away.

"There go your falsehoods," the rabbi says to his visitor. "Go get them, and bring them back."

When we work with data, we are working with information that is uniquely powerful. It reflects a human imperative to come to grips with our

G. Langer (✉)
Langer Research Associates, New York, USA
e-mail: glanger@langerresearch.com

D.L. Vannette, J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, https://doi.org/10.1007/978-3-319-54395-6_45

surroundings, to understand our communities, our societies, our nation and our world through the act of quantification. You don't have to read the Book of Numbers, or to know the meaning of the writing on the wall – "*mene, mene, tekel, uparsin*" – "numbered, numbered, weighed and divided" – to recognize the singular authority with which data speak. This reality requires us to be particularly prudent and careful in producing and delivering data to the world.

Good data are powerful and compelling. They lift us above anecdote, sustain precision and expand our knowledge, fundamentally enriching our understanding and informing our judgment. They're absolutely essential. Other data, by contrast, may be manufactured to promote a product or point of view. Intentional manipulation aside, they can be produced unreliably using suboptimal techniques or inadequate methodologies.

Some of these methods leave the house of inferential statistics. As such they lack a firm basis for the validity and reliability that we seek. The very low bar of entrance to these alternative methods brings in nonprofessional participants, untrained in sampling principles, questionnaire design, and data analysis. The result can be numbers and percentage signs that would seem to speak with authority, but that in fact can misinform or even be used to disinform our judgment.

We need to continue to subject established methods to rigorous and ongoing evaluation, but also to subject new methods to these same standards. To value theoreticism as much as empiricism – they are equally important – to consider fitness for purpose, and fundamentally to go with the facts, not the fashion.

A brief background on survey sampling may be of use. We start with the full population, a census survey. It can give you a very high level of accuracy, but it's prohibitively expensive and highly challenging to produce, as our research colleagues at the Census Bureau can confirm.

The first alternative is availability or "convenience" sampling, straw polls for example. They're quick and inexpensive. The risk is that they don't have a scientific basis or theoretical justification for generalizing to the full population. The pushback from elements of the research community is "*Hey, they seem to work – they're good enough,*" or in today's much-abused phrase, they are "*fit for purpose.*" That argument is not new; *The Literary Digest* did straw polls for decades, from 1916 to 1932, and correctly predicted the presidential election in each of them.

Then the 1936 election rolled around. The Literary Digest sent out 10 million cards to magazine subscribers, addresses from phone books and automobile registration lists. They got two-and-a-half million back and they found an easy win for the wrong candidate.

What happened? One, coverage bias: The sampling frame differed in important ways from the population of interest. Republicans were more likely to have been in the sample – skewing upscale, they were more apt to be able to afford subscriptions, telephones, and cars during the Great Depression. Next, we had systematic survey response, with Republicans more likely to participate, perhaps in order to express discontent with the incumbent Democratic president and his New Deal policies.

Given this failure, after 1936, availability sampling largely was replaced with quota sampling, which had correctly predicted FDR's win that year. Quota sampling attempts to build a miniature of the population. It identifies what are supposed to be key demographic variables; researchers then go out and seek to mirror those groups in their samples. That's said to produce representative estimates.

How did that work out? Not great, as Thomas Dewey could attest. There are a few reasons for the polling debacle of 1948. One was timing. Pollsters stopped collecting their data early, assuming things wouldn't change – when they did. Pollsters erroneously assumed that undecided voters would break as decided voters had; that's unsupported. Likely voters also may have been misidentified. But another problem involved sampling. Purposive selection appears to have allowed interviewers to choose more-educated and better-off respondents within their assigned quotas – pro-Dewey groups. This marks a key risk in purposive sampling – the risk of unintentional (or, indeed, intentional) systematic bias in the selection of individual survey respondents.

Even if quota sampling weren't entirely to blame, the Dewey-Truman fiasco highlighted inherent limitations in the method beyond those of respondent selection. One is that it's impossible to anticipate, let alone match, every potentially important demographic group. If a researcher wishes to match on gender, Hispanic ethnicity, race, education, age, region, and income within customary categories, that yields 9,600 cells raising the question of how many interviews can feasibly be done, and whether it's enough to fill those many cells. Further, as with availability sampling, quota sampling lacks theoretical justification for drawing inferences about the broader population.

The research community had this out in the early, formative days of the American Association for Public Opinion Research (AAPOR). Probability sampling – a random sample drawn with known probability of selection – won out. As Hanson and Hauser said in *Public Opinion Quarterly* in 1945, "an essential feature of reliable sampling is that each element of the population being sampled…has a chance of being included in the sample and, moreover, that that chance or probability is known."

The principle behind this thinking in fact goes back a little further, to the philosopher Marcus Tullius Cicero in 45 B.C. Kruskall and Mosteller quoted him in 1979 and I do so again here:

> Diagoras, surnamed the Atheist, once paid a visit to Samothrace, and a friend of his addressed him thus: "You believe that the gods have no interest in human welfare. Please observe these countless painted tablets; they show how many persons have withstood the rage of the tempest and safely reached the haven because they made vows to the gods." "Quite so,' Diagoras answered. 'But where are the tablets of those who suffered shipwreck and perished in the deep?"

Cicero's lesson on the perils of non-probability sampling have echoed throughout the ages, expressed in modern times, for example, by George Snedecor in the *Journal of Farm Economics* in 1939, Johnson and Jackson's *Modern Statistical Methods* in 1959, and Leslie Kish's *Survey Sampling* in 1965. There are many such cites. What they share in common is an expression of the fundamental theoretical principles of inferential statistics, which tell us clearly how and why probability sampling works.

Non-probability sampling has yet to enunciate any such operating principle. Yet that doesn't seem to deter its practitioners. Non-probability based Internet opt-in surveys are estimated to be a $5–$6 billion business in this country. A vast amount of market research has moved into Internet opt-in samples, and much else is done there as well. Let's explore what we know about this work.

My personal journey of discovery started 15 or so years ago as I set up the first standards and vetting operation for survey research by a national news organization, at ABC Network News. With the support of management, we put in place a system in which any survey research being considered for reporting would come to my group first; we would check it out, and either clear it or kill it.

The sort of things we saw include, for example, a report in my local paper of a poll, picked up from the *Sunday Times* of London. ABC was interested. We looked up the *Sunday Times* and indeed there it was, a poll of nearly 2,000 people by an outfit called YouGov. I was unfamiliar with them at the time this was 2003. We checked out their website: "Voice your opinion and get paid for it," it said. "Register now."

We looked further around the Internet and found a lot of photos of people waving their arms in the air, really excited about the money they are earning taking surveys on the Internet. In effect, it turns out, these are poll-taking clubs comprised of individuals who have signed up to click through online questionnaires in exchange for points redeemable for cash and gifts.

The missives also go out by e-mail. Here's one a friend of mine at Princeton received, and I suspect that every college student in America has seen it: "We have compiled and researched hundreds of research companies that are willing to pay you between $5 and $75 per hour simply to answer an online survey in the peacefulness of your own home." Or skip the cash: another we found raffled a car.

You might imagine the attraction of signing up for these things in a variety guises to increase your chances of getting selected to take surveys to win cash and gifts. Just as a test, a colleague of mine signed up for one of these online panels. He identified himself as a 32-year-old, Spanish-speaking, female, African-American physician, residing in Billings, Montana, and received surveys to fill out starting that very same week.

Controls are possible. If you've got to sign up with an address, and then you have to have your checks and gifts sent to that address, there could be some match there, an identity check. Rather, we find redemption pages with instructions that take a very different direction, suggesting that you can have your reward sent to a friend. Collecting prizes for a whole bunch of alternative personalities seems a lot easier.

Does this happen? Do people actually burn through many, many surveys if they're interested in racking up points redeemable for cash and gifts? One report found that among the 10 largest opt-in panels, 10 percent of participants accounted for 81 percent of survey responses, and indeed 1 percent of participants accounted for 24 percent of responses.

Questions are apparent. Who joins these poll-taking clubs? What verification and validation of respondent identities is undertaken? What logic and quality control checks are put in place? What weights are applied, on what theoretical basis, from what empirical source, and to what effect? What level of disclosure is provided, just for example in terms of the use of survey routers? What claims are made about the quality and qualities of the data, and how are they justified?

These are important questions if we are going to try to understand these data and try to come to some judgment about the approach. Such questions should be asked about all survey research – probability-based, and non-probability alike. We need to ask them so we can assess claims like this one, from Google Consumer Surveys, saying that its samples "produce results that are as accurate as probability-based panels." This claim is made in relation to a product in which you can buy the ability to ask one or two questions delivered as pop-ups to users of a search engine who are looking at something called "premium content." The demographic data that then is associated with the responses to these one or two questions apparently are imputed through analysis of users' IP addresses and previous page views.

Interestingly, you can go to Google and see who they think you are on the basis of your IP address and previous page views. A young woman on my staff was identified as a 55-year-old male with an interest in beauty and fitness. Another as a 65-year-old woman, double her actual age, with a previously unknown interest in motorcycles. And I was identified by the Google imputation as a senior citizen, thank you very much, in the Pacific Northwest, which I've visited exactly four times in my life, with an interest in space technology, which is news to me.

Another non-probability data provider lays claim to Bayesian "credibility intervals." This looks like no more or less than the typical formula to compute a margin of sampling error for a probability sample. Indeed, by whatever name, claims of a margin of sampling error associated with convenience samples are commonplace, without seeming justification.

What's the rationale for all of this? Per one provider's web page: "Traditional phone-based survey techniques suffer from deteriorating response rates and escalating costs." So, the pitch goes, we've got something else going instead. This is more a knockdown than a build-up argument: The old stuff costs too much and has got low response rates, so let's just throw in the towel on supportable inference.

But let's dissect this a little bit. We cannot achieve perfect probability; there are no 100 percent response-rate surveys. The Pew Research Center, as Scott Keeter reports in this volume, has observed a dramatic decline in its response rates, from 36 percent to 9 percent. (Some of us do better.)

Does this trend of declining response rates poison the well? Extensive research consistently has found that response rates in and of themselves are not a good indicator of data quality. There is theoretical support for this empirical result. If non-response to surveys itself is largely a random rather than a systematic phenomenon in terms of the variables of interest, then it does no substantive damage to inference. And there are a lot of carefully produced data to support that conclusion.

Response rates were one knockdown in the pitch I showed you. The other was that costs are escalating and it is cheaper to use an opt-in panel. No argument; some vendors have offered them at $4.00 per complete. That's a tenth or less of the cost of high-quality probability-sample research. The next step then is to ask, is it worth it? To know, we have to move to empirical testing of these data.

Yeager and his colleagues wrote an important paper in 2011 comparing seven opt-in online convenience sample surveys with two probability-sample surveys. The probability sample surveys were consistently highly accurate; the online surveys were always less accurate and less consistent in their level of accuracy. Results of the probability samples were less apt than the convenience samples to be significantly different from benchmarks, and the highest single error was dramatically different.

Opt-in online survey producers like to compare their results to election outcomes: the claim is that if you do a good pre-election estimate of a political contest, therefore you've got good data. I suggest that this is an entirely inappropriate basis for comparison. Pre-election polls are conducted among an unknown population; we don't know who's going to vote, so we resort to modeling and perhaps weighting for likely voters. That introduces judgment, often applied in an opaque fashion. In a good estimate, are we seeing reliable polling, or just good modeling?

Administrative benchmarks against which we can measure unmanipulated survey data offer a far more reasonable basis for comparison. In the Yeager et al. study, you can see the average absolute errors across opt-in panels compared to the probability samples, and the sizable inconsistencies across those non-probability sources. The average absolute errors were significantly different; the largest absolute error as well.

This paper also found little support for the claim that some non-probability panels are consistently more accurate than others – so you really don't know if you've landed on a good panel or not, or if you've got a good panel, whether it will be good next time. Weighting did not always improve the accuracy of the opt-in samples. There was no support for the idea that higher completion rates produce greater accuracy. And the probability samples were not just more accurate overall, but also more consistently accurate.

This study suggested that it's virtually impossible to anticipate whether an opt-in survey will be somewhat less accurate or substantially less accurate, or whether knowing that it is accurate on one benchmark can tell you whether or not it will be accurate on others. Yeager and his coauthors said this shouldn't be a surprise because there is no theory behind it. And they warned that you can cherry-pick results to create the appearance of reliability and validity in these data when in fact it may not be there on a systematic basis.

Other studies reach similar conclusions. The Advertising Research Foundation produced a study, "Foundations of Quality"; one element that was released found far more variation in estimates of smoking prevalence across opt-in online panels than in probability methods. As one commentator put it, “The results we get for any given study are highly dependent (and mostly unpredictable) on the panel we use. This is not good news.”

Mario Callegaro, at the 2012 national AAPOR conference, presented a review of 45 studies comparing the quality of data collected via the opt-in online method versus either benchmarks or other modes. Findings were similar to those we saw from Yeager, Krosnick. In summary:

- Panel estimates substantially deviated from benchmarks, and to a far greater degree than probability samples
- High variability in data from different opt-in panels
- High levels of multiple-panel membership (between 19 and 45 percent of respondents belong to 5+ panels)
- Substantial differences among low- and higher-membership respondents
- Weighting did not correct variations in the data

Opt-in Internet polls aren’t the only non-probability samples of concern. Mike Traugott, of the University of Michigan, gave a presentation at AAPOR in which he compared the accuracy of four low-cost data collection methods. One was Mechanical Turk, a panel comprised of people who sign up on the Internet to participate in studies in exchange for minimal compensation. Another was an IVR or robo-poll. The results indicated that the unweighted demographic outcomes were substantially different from benchmark data. Mechanical Turk, for example, is a good place to find Democrats, but it’s probably not a good place to find a good sample. And whether it was unweighted or weighted to standard demographic variables didn’t help.

Additionally, we have AAPOR’s report on online panels, produced in April 2010. It said that researchers should avoid non-probability online panels when one of the research objectives is to accurately estimate population values. This report talked about the underlying principles of theory. It talked about the non-ignorable differences in who joins these panels. It also said that the reporting of a margin of sampling error associated with an opt-in sample is misleading. In sum, AAPOR concluded: “There currently is no generally accepted theoretical basis from which to claim that survey results using samples from nonprobability online panels are projectable to the general population.” In 2013, AAPOR produced an additional report on non-probability sampling in general; while encouraging additional research and experimentation, it also noted the absence of a theoretical framework to support inference, among other challenges.

Apart from population values, AAPOR's report on opt-in Internet panels left room for this: Perhaps we can still use convenience samples to evaluate relationships among variables and trends over time. In 2010 Pasek and Krosnick found otherwise. They compared otherwise identical opt-in online and RDD surveys sponsored by the U.S. Census Bureau assessing intent to fill out the census, produced in order to inform a marketing campaign to encourage participation. Many of the findings of this research replicate what we saw previously with the Yeager et al. study, including that the telephone samples were more demographically representative than the opt-in Internet surveys, even after post-stratification. Pasek and Krosnick also reported significantly, substantively different attitudes and behaviors across the two data streams.

What was new was that they also reported instances in which the two data streams, as they put it, "told very different stories about change over time," as well as suggesting different predictors of intent to complete the census – the basic information you want and need if you're going to create a campaign to improve census compliance. Who is not going to complete the census? The authors indicate that the Census Bureau would have arrived at fundamentally different conclusions from these two different data sources.

So we have very different research conclusions from these two datasets. One is based on a probability sample, the other on a non-probability source. The notion that, regardless of estimated population values, you can use non-probability samples to draw conclusions about correlations in data is now in question. More research is needed. But the picture this research paints is not pretty.

Can non-probability samples be fixed? Bayesian analysis is all the buzz; the question is whether it's simply the magician's cloak. What variables are used, how are they derived, are we weighting to the dependent variable or to its strong expected correlates, and is that justified? Sample balancing is another suggestion – in effect time-traveling back to quota sampling. Can we take this non-probability sample get our 9,600 cells filled up and be good with it? Let's return to Gilbert et al. back in 1977: "The microcosm idea will rarely work in a complicated social problem because we always have additional variables that may have important consequences for the outcome."

Some say otherwise. A recent paper on political behavior, published in an academic journal, peer reviewed, based on non-probability sample YouGov/Polimetrix data, claimed to use sample-matching techniques to build representative samples with quality that "meets, and sometimes exceeds" that of probability sampling.

AAPOR'S task force on opt-in panels, as we have seen, says otherwise. And AAPOR has company. Among many others, Paul Biemer and Lars Lyberg have offered similar conclusions; to quote from their book

Introduction to Survey Quality: "Unfortunately, convenience samples are often used inappropriately as the basis for inference to some larger population." Further, they state, "…unlike random samples, purposive samples contain no information on how close the sample estimate is to the true value of the population parameter." These references are not intended to stall further research, but rather to encourage its sober practice. Close and continued evaluation of these methodologies is critical. So are full, forthright disclosure, and defensible claims about the data's validity and reliability.

We also need to evaluate research stemming from social media. This is a giant mass of data – truckloads of it, generated on a minute-by-minute basis. There is a lot of work underway trying to grapple with how to understand it. I appreciate, value, and encourage that work, but I also suggest that it should be held to reasonable research standards and scrutiny.

Consider the sampling challenges. We may want to assume that a tweet or a Facebook post represents one individual expressing his or her actual opinion on something once. In fact, users are not limited; some can post incessantly and others rarely. Users can have multiple accounts. Some accounts don't belong to individuals at all, but rather to companies, organizations or associations that use them, not to express individual attitudes, but to promote products or points of view. These posts can be driven by public relations campaigns including paid agents using automated bots. And this all exists within this giant data blob.

Parsing it out is highly challenging. Assuming users are people, their location often is not accurately captured, if at all. Users, of course, are self-selected, so again we're lacking a theoretical basis to assume that this is somehow representative of others. Recent research finds that among all online adults – and again there is still a substantial population that is not online at all – 15 percent are Twitter users at all, 8 percent daily. It's been estimated that fewer than 30 percent are Americans – a challenge if you're purporting to measure, say, U.S. election preferences – and of course there is a substantial age skew.

Let's skip over the sampling issues and say we're going to take this stuff anyway and figure out what it means. Determining the meaning of posts or tweets requires content analysis. Traditionally, we would do this through independent, human coders, with a codebook and a measure of inter-coder reliability. But that sort of approach is unrealistic given this volume of data. So researchers now generally use computerized analysis programs. They're lower cost and they're faster, but problems arise. Tweets are chock full of slang, irony, sarcasm, abbreviation, acronyms, emoticons, contextual meaning, hashtags – a highly complex dataset from which to derive meaning. There also is ambiguity in determining target words.

Say we are going to grab every tweet posted during the 2012 presidential debates that used the word "Obama." But what about tweets that used the word "the president" or "the pres" or "Barack" or "BO" or other phrases meant to refer to the president during the debate? We may miss those. And even if we figure out some way through this, we still don't have contextual data, demographic or other attitudinal data that will help us to reach intelligent research conclusions.

Kim et al. gave a presentation at AAPOR in 2012 in which they compared computerized content analysis systems to human coding in the ability to code Twitter comments as positive, negative, and neutral. The good news is that they weren't dreadful on neutral tweets in terms of matching automated coding to human coding. But the bad news is that on the positive and negative tweets, which we're likely most often interested in, they were dramatically different. This inconsistency is a cause for concern.

Rather than attempting to parse the content, some researchers are trying to focus just on the sheer quantity of posts about a certain topic. One study found that in the German national election, the proportion of tweets mentioning a political party in fact reflected its vote share. But it also found an enormous skew in the creation of these items – 4 percent of users accounted for 40 percent of the messages. This seems to be vulnerable to an orchestrated campaign to promote a political party. Similar analyses in the United States have differed; one, for example, found that Google Trends data did worse than chance at predicting election outcomes.

Twitter, again in 2012, posted something called the Twitter Political Index. Just what it was or how it worked was not disclosed. It was scaled on 0–100, but apparently not a percentage. You're Mitt Romney, you're Barack Obama, you get a number. And the numbers varied widely. On August 6th, Obama's number was 61. On August 7th, a day later, Obama's number was 35. Publicly released probability-sample surveys showed no volatility of this type. The meaning of this exercise, its purpose and its contribution to our understanding of the election were unclear, to say the least.

It's been suggested that Facebook user groups can be used for snowball sampling, to try to build a sample of hard-to-get respondents. A paper by Bhutta in 2011 reported on this approach to reach Catholics. The author reported that it was faster and cheaper than traditional methods, but there were vast differences in the Facebook snowball sample that she obtained versus General Social Survey data in terms of demographic characteristics of Catholics, including gender, race, education, and mass attendance. Bhutta concluded that the Facebook respondents could not possibly serve as a representative sample of the general Catholic population.

We need to touch, as well, on falsification. It was suggested in 2012 that as many as 30 percent of Obama's Twitter followers and 22 percent of Romney's were fabricated. Forbes magazine reported how individuals can purchase 25,000 fake Twitter followers for $247, and that these bots automatically will tweet snippets of text. Are we going to use this information to assess public attitudes?

What does the future hold? In probability sampling, we need continued studies of response-rate effects; whether it's 19 or 9 percent, it's not what it used to be. Efforts to maintain response rates can't hurt. Renewed focus on other data-quality issues is important including quality control, coverage, and questionnaire design, too often the forgotten stepchild of survey research. Equally important is the development of probability-based alternatives, including online and mobile-based administration and improved panel-management techniques.

In non-probability sampling, further study of the appropriate as well as inappropriate uses of convenience sample data are very much in order. Further evaluation of well-disclosed and emerging techniques will be important; this means enhanced disclosure and a more sober approach than we often see today.

In social media, a vast and growing source of data, we also need further evaluation of the appropriate use of this material. We can study, for example, whether and how social media relates to attitude formation and change in public opinion, perhaps as a compliment to reliable attitudinal measurement. The key in all cases is to establish the research question, evaluate the limitations of the tools available to answer that question, and then go forward with informed judgment, full disclosure, and honest assessment on the basis of empirical results and theoretical principles alike.

Gary Langer is a survey research practitioner. He is president of Langer Research Associates and former long-time director of polling at ABC Network News. Langer is a member of the Board of Directors of the Roper Center for Public Opinion Research, a trustee of the National Council of Public Polls and former president of the New York Chapter of the American Association for Public Opinion Research. His work has been recognized with 2 News Emmy awards, 10 Emmy nominations, and AAPOR's Policy Impact Award. Langer has written and lectured widely on the measurement and meaning of public opinion.