Understanding Polling Methodology

BY **MATTHEW MENDELSOHN** AND **JASON BRENT**

RÉSUMÉ ► La recherche par sondages est un outil de recherche dont l'importance est sans cesse croissante auprès des agents publics, des politiciens et des universitaires. Bien qu'on ait observé des progrès dans le domaine scientifique des sondages, il y a aujourd'hui plus que jamais des sondages mal exécutés. Comprendre les objectifs et la méthodologie des sondages n'est pas toujours facile : comment est-il possible de différencier un bon sondage d'un mauvais ? Comment faut-il poser les questions qui permettront de mieux connaître l'opinion publique ? Enfin, est-il possible d'évaluer cette opinion en posant un petit nombre de questions à un minuscule pourcentage de la population visée ? Les sondages peuvent procurer aux décideurs de précieux renseignements, mais savoir comment les concevoir et les utiliser est primordial pour en tirer le maximum. (Traduction : **www.isuma.net**)

ABSTRACT ► Survey research has become an increasingly important tool for public servants, politicians and academics. While there have been advances in the science of polling, there are now more poorly executed polls than ever. Understanding the purpose and methodology of polling is not always easy: How can we distinguish a good poll from a bad one? How can we ask questions to better understand real public opinion? Is it possible to assess public opinion by a small number of questions asked to a tiny percentage of the entire population? Polls can provide great insight for decision makers, but understanding how to design and use them is critical to maximizing their usefulness.

ILLUSTRATIONS : NORMAND COUSINEAU

THE PROCESS OF POLLING is often mysterious, particularly to those who don't see how the views of 1,000 people can represent an entire population. It is first and foremost crucial to remember that polls are *not* trying to reflect individuals' thinking in all their complexity; all they are trying to do is estimate how the population would respond to a series of closed-ended options, such as, "should tax cuts or new spending be a higher priority?"

The sampling process

The basic principle of polling is that a sample of a population can represent the whole of that population if the sample is sufficiently large (the larger the sample, the more likely it is to accurately represent the population) and is generated through a method that ensures that the sample is random. The random sample—one in which everyone in the target population has an equal probability of being selected—is the basis of probability sampling and the fundamental principle for survey research. Beyond a certain minimum number, the actual number of respondents interviewed is less important to the accuracy of a poll than the process by which a random probability sample is generated.

Most polls conducted today use telephone interviewing. Interviewers are not usually permitted to vary the question in any way, and no extemporaneous explanations are allowed. In-home interviews can also be conducted but these are more expensive. Surveys can be conducted by mail or on-line, though there are concerns that samples will not be representative (although for targeted populations with an incentive to reply, these methods can be efficient and cost effective). For telephone surveys, the usual method is *random digit dialing (RDD)* or use of the most recently published telephone listings. If using the latter method, one usually chooses telephone numbers randomly, but then changes the last digit to ensure that unlisted numbers and those who recently got telephones have an equal chance of being selected.

Cluster sampling is usually used for in-home interviewing, whereby representative clusters in particular neighbourhoods are selected (the costs of travelling hundreds of kilometres to conduct one interview make a truly random sample impossible for in-home interviews). Quota sampling can also be used. With this method, the percentage of the population that falls into a given group is known (for example, women represent 52% of the population) and therefore once the quota is filled (520 women in a sample of 1000) one ceases to interview women, instead filling one's male quota. Quota sampling is used less frequently because of the simplicity of RDD techniques, and because of the possibility to weight the data afterwards. Weighting the data is a process that ensures that the sample is numerically representative of various groups in the population. For example, to be able to make reasonable extrapolations about, say, Atlantic Canadians, in a survey, it may be necessary to "oversample" Atlantic Canada. However, when looking at the national results, each Atlantic Canadian respondent would be counted as less than a full respondent, otherwise Atlantic Canadians would be overrepresented in the sample. This is what "weighting the data" means, and it is customary to weight the data by region, gender and age to ensure representivity.

Polling firms now tend to use the CATI system (computer assisted telephone interviewing) in which the interview process is streamlined. Answers are automatically submitted into a data bank, question filters can be used that ask different respondents different questions depending on their previous answers, and the wording of questions can be randomly altered.

Understanding error: margin and otherwise

We have all heard poll results in news stories described as being, "accurate to plus or minus three percentage points, 19 times out of 20," but what does this mean? This statement, and the figures that it contains, refers to the "sampling error" (3%) and "confidence interval" (95%, or 19 out of 20) of the poll that has been taken. This means that 95 percent of all samples taken from the same population using the same question at the same time will be +/- the sampling error (usually referred to as the "margin of error").[1]

The reported margin of error assumes two things: that the sample was properly collected (and therefore represents the cross-section of the target population), and that the questions are properly designed and in fact measure the underlying concepts that are of interest. If either of these two assumptions is violated, the margin of error underestimates the real difference between the survey results and what the population "thinks." The reporting of the margin of error may therefore create an aura of precision about the results that is in fact lacking. If the sample is not truly random, or the questions are poorly worded, then the margin of error underestimates the degree of uncertainty in the results.

If the first assumption is violated—and the sample is poorly collected—one has a biased sample and the survey results could be a poor estimate of what the population actually thinks, above and beyond the margin of error. Even when well done, surveys may be somewhat biased by under- or over-representing certain kinds of respondents. For example, telephone polls obviously under-represent the very poor (those without telephones) and the homeless; those who do not speak English or French well may be under-represented if the polling firm does not have multilingual interviewers. An additional concern is the low response rate (the percentage of people who are actually reached that agree to answer questions). It is possible that the individuals who agree to participate in the survey (as low as 20% now in many telephone polls) may not be representative of those who do not participate.

If the second assumption is violated—and the questions are poorly designed—one has measurement error. A survey is no more than a measurement instrument. Like a bathroom scale which measures the weight of a person, a survey question measures people's attitudes. In the world of the physical sciences, we can develop scales that are almost perfectly accurate, but in the world of attitude measurement, our instruments are far less perfect, in part because so many of the underlying concepts are socially and politically contested: what is the "right" question to ask to measure opinion on increasing health care spending? There isn't one right question. Dozens of equally credible questions could be posed, each producing somewhat—or even wildly— different results, and each only an approximation of the "real attitudes in people's heads" toward increasing health care spending.

Any consumer of polls must also be aware that some coding error is inevitable (the interviewer punches "1" on the keypad instead of "2"). Although this doesn't seem to be too serious, on rare occasions whole questions have been miscoded—with all of the "1's" and "2's" being inadvertently mislabeled. Extremely odd or deviant results should be checked with the polling firm for possible coding error.

Another widespread and challenging problem is reporting error (as occurs when respondents lie, fail to remember accurately, or offer what they deem to be a socially desirable response). Reporting error can either result from individuals (some people are simply more likely to misreport) or from the item (some questions, concerning income or sexual behaviour for instance, are more likely to produce false reports). Companies that monitor TV viewing, for example, found that when they went from the use of self-reporting to actual electronic monitoring, PBS viewing dropped dramatically, while more people were now watching wrestling.

Internal or external validity checks can be used to reduce error as much as possible. For internal checks (i.e., internal to the survey) a number of different but similar questions can be asked at various points during the survey. Or the sample can be split, with half the sample being asked one version of a question and the other half another, slightly different version of the question. If results are fairly similar from question to question, one can be fairly certain of internal validity. External validity checks include comparing the survey results to other data sources. This is often impossible, but sometimes (using census data, for example) it is quite easy to verify the accuracy of the results. In the case of assessing vote intentions, there is always, of course, the "external validity check" provided by election results.

What can you measure?

We usually think of polls as "opinion polls," but "opinions" are only one of four measurables. Polls can also measure behaviour, knowledge and socio-demographic characteristics. For these items, the questions themselves are often easier to formulate and less prone to debate. To find out whether respondents voted, know about changes to a government program, or whether they were born in Canada, the question is—usually—fairly straightforward; how to construct an opinion question properly, on the other hand, is often highly contentious.

Opinions themselves can either be what social scientists call "opinions," "attitudes" or "values." Values can be understood as basic beliefs held by individuals which remain relatively immune to change and which play an important role in individuals' lives and choices. Opinions are judgments about contemporary issues and policy options. Attitudes represent an intermediate category between values and opinions: they tend to be fairly well formed and settled world views that can be used to assess new issues. The measurement of values and attitudes is often more useful because these represent more enduring views, rather than the ephemeral opinions that may be heavily influenced by short-term events. But most polls of relevance to policy makers deal with opinions regarding current policies, such as views on a flat tax or same-sex benefits, even though these opinions are heavily influenced by underlying attitudes and values toward the free market and moral traditionalism.

Reporting of the margin of error may create an aura of precision about the results that is in fact lacking.

The measurement of opinions, attitudes and values is complicated because there are many different dimensions of an attitude. Does it actually exist—does the person have a belief about the question? What is the direction of this attitude—support or oppose, yes or no, etc. What is the intensity of this view—someone who strongly agrees with something is quite different from someone who has only moderate feelings about an issue. Although existence, direction and intensity are key, one could also try to measure the level of certainty the respondent expresses, how well informed the attitude is, and whether it is important to the respondent. Many of these can be thought of as "opinion crystallization": how well formed is the person's view? All of these can be measured, and all should be considered when drawing conclusions about survey data. It is therefore often useful to measure first whether an opinion exists, then its direction, then its intensity, then the importance or salience of the issue to the respondent. To understand how public opinion may affect policy, it is important to know how intense opinion is. For example, an intense minority that mobilizes supporters may have a greater impact on public policy than a disinterested and ambivalent majority.

One must always be on guard for non-attitudes, the expression of an opinion when one does not really exist. Non-attitudes are generally considered to be random, and are heavily influenced by context—cues in the question or what was on TV last night. One risks measuring non-attitudes if one has failed to properly measure the first dimension of the attitude: does it actually exist? The reporting of non-attitudes can be minimized by making it easy for respondents to say that they "don't know" by including softening language in the question, such as "some people may not have heard enough about this issue yet to have an opinion." But some survey organizations are reluctant to take "I don't know" as an answer from respondents because it reduces sample size, sometimes drastically on obscure issues. In essence, there is a financial incentive to avoid "don't knows." Some surveys will discourage "don't knows" by not presenting it as a possibility or by asking undecided respondents a follow-up question to probe which way they are leaning. The risk inherent in this approach is that some respondents who have no opinion will feel pressured to respond, expressing a non-attitude. When respondents are explicitly offered a painless opportunity to say that they have not thought enough about the issue, the number who say they have no opinion significantly increases.

"Opinions" are only one of four measurables. Polls can also measure behaviour, knowledge and socio-demographic characteristics.

The wording and format of questions: the core of questionnaire design

Public opinion cannot be understood by using only a single question asked at a single moment. It is necessary to measure public opinion along several different dimensions, to review results based on a variety of different wordings, and to verify findings on the basis of repetition. Any one result is filled with potential error and represents one possible estimation of the state of public opinion. The most credible results emerging from polls are usually those which either examine change over time or rely on multiple indicators (questions) to get a better understanding of the phenomenon in question. Examining the relationship between questions is also exceptionally useful, and it is important not to think of these results as absolutes (e.g., "Men are 20 points more likely than women to believe X.") but as general tendencies ("There is a strong relationship between believing X and being male.")

A polling question should be thought of as a measurement tool. We presume that something—opinions, behaviour, etc.—exists out there and we want to measure it. The best way we can measure it is by asking questions. Question wording will inevitably heavily influence results. The effect of question wording can sometimes appear idiosyncratic: as new issues arise, it is difficult to know the precise effects of question wording without testing 50 other possible wordings. For example, polling questions over the last few years which have tried to measure respondents' views on whether tax cuts, new spending or debt reduction should be a higher priority for governments have been fraught with measurement error. It turns out that phrasing the spending option as "new spending on social programs" received only modest support (about 33% of Canadians preferred this option, according to an Environics/CROP survey conducted in 2000 for the Centre for Research and Information on Canada), while "putting money back into health care and education" received far more support (the choice of 45% of Canadians, according to an Ekos survey asked at about the same time). Neither option is necessarily "the right way" to ask the question—in fact, asking the question in both ways and finding such different responses actually tells you a great deal about what people are thinking in terms of priorities for the surplus. But such knowledge is the result of trial and error and/or well-designed experiments where the wording of questions is varied from respondent to respondent, not any methodological rule.

Changes in the format of questions can have a more predictable, less idiosyncratic impact on results. For example, questions can be either open or close-ended. Closed-ended questions, which provide respondents with a fixed set of alternatives from which to choose, are used far more frequently in surveys because they are easier to code and less expensive to collect. Open-ended questions ask respondents to offer an opinion or suggestion, or answer a question without predefined categories for the response. Although open-ended questions are infrequently used, they can serve as a useful first step in identifying which closed-ended items to use in a list. Open-ended questions are, however, used more often than one might think, sometimes with a dramatic effect on results. For example, the simple and standard vote intention question can be affected by whether the respondent is provided with a list of alternatives. When a new party is in the process of formation, the inclusion of the party as an option in a survey may prompt respondents to recall that the party exists, and remind them that they have a favourable impression of the party. In the years following the 1988 federal election, when the Reform Party had begun to make an impact but had yet to elect any members during a general election, the Reform Party fared poorly with a closed-ended question that provided a list of the traditional parties that excluded Reform; Reform tended to fare a bit better if the question was open-ended because Reform voters would not have to offer a response that was not offered in a closed list; and Reform fared best when the survey added Reform to the closed list.

The inclusion of a middle position in a question can also seriously affect a poll's results. Significant increases in the number who choose the status quo position or middle of the road position are found when such a response category is explicitly offered. Some researchers assume that individuals who choose the middle option actually prefer one of the two directional positions, even if it is with little intensity, and would make a choice if forced to do so. Others assume that the middle position is a valid choice reflecting real attitudes. There is some evidence that including a middle position will attract those respondents who really have no opinion, and who should really be counted as undecided. If respondents are asked whether they think the government should be spending more, less or about the same amount of money on a particular public policy issue, many who say "the same" may in fact have no opinion. Including a middle position may therefore overestimate the number who are in the middle position and hence create a bias toward the status quo on many issues. On the other hand, by not offering a middle position, one may create a false impression that opinion is polarized, when in fact many people may be somewhere in the middle and genuinely satisfied with the status quo.

Asking respondents to agree or disagree with a series of statements is often a cost-effective and rapid way to ask a large number of questions. However, these types of surveys are also highly problematic because, for a variety of reasons, respondents are more likely to agree than disagree. This phenomenon is referred to as an "agree-response bias," or the "acquiescence effect." Some people may simply be psychologically predisposed to agreeing and acquiescing to the interviewer. More important, acquiescence can be the product of the one-sided nature of the statement or the lack of any perceived alternative. It is quite easy to formulate a statement on most sides of an issue with which most respondents will agree when it is divorced from real-world trade-offs or alternatives. One could get strong levels of agreement with both the statements: "I would like my taxes cut" and "The government should invest more in the health care system," without these responses offering any meaningful guidance to governments interested in reflecting public opinion in policy priorities. Because of the acquiescence effect, it is often beneficial, when possible, to make sure that questions are balanced by forcing respondents to choose between two conflicting statements, both of which are written to be as appealing as possible.

To understand how public opinion may affect policy, it is important to know how intense opinion is.

Once it has been established that the respondent has an opinion and the direction of that opinion is known (favour or oppose; agree or disagree, etc.), it is important to ascertain how intensely that opinion is held. Likert scales are the commonly used form for measuring intensity. A Likert scale often runs from "strongly approve" to "approve" (or "somewhat approve") to "disapprove" (or "somewhat disapprove") to "strongly disapprove." There may also be a middle category to measure neutrality. It is important to be aware that survey organizations often collapse the two "agree" and two "disagree" categories together when reporting results. If most people fall into the "somewhat" categories, it is clear that there is less crystallization on the issue in question, and a different public opinion environment exists than if most respondents situate themselves in the "strongly" or "very" categories.

In addition to Likert scales, numerical scales are often used. One can use a 5, 7, 9 or 10-point scale to measure agreement or disagreement. For these scales to be easily interpreted, the question must make it clear where the neutral point is so that respondents can anchor their responses. Thermometer scales (0-100) are often used to measure how warmly respondents feel toward individuals, groups or objects. The practice of collapsing respondents together into general agree/disagree categories may be problematic because (for example, using a 7-point scale), those who score 5 are grouped with those who score 7, when those who score 5 may in fact be quite neutral and have more in common with those who score 4. This does not mean there is anything wrong with the use of a 7-point scale, but it does mean that results should be read carefully.

For questions to be effective, the number of choices offered to respondents must be balanced. There should be the same number of categories on the scale representing both directions of opinion for the given question. If there is an even number of categories, there should be no neutral point; if there is an odd number of categories, there should be an anchored, neutral mid-point.

It is difficult for respondents to remember complex or lengthy lists during a telephone interview. Therefore, questions should not have more than three or, at the most, four categories, though there are exceptions to this rule. For example, on a vote intention question or where respondents are asked to identify their level of education, a greater number of categories is acceptable. It is also possible to offer a greater number of choices when there is an implicit order to the answers.

Increasingly, surveys are being used by politicians, parties and public institutions to test messaging. In such situations, many of the above rules are thrown out the window and questions can be double-barrelled or loaded because what one is interested in is a general reaction to a statement, not a firm conclusion about what percentage of the population supports a given policy. Is the statement appealing or offensive to respondents? Which of two statements is more appealing? In such situations, question wording experiments are particularly popular. Some of the words in the question can be systematically varied, and the results can be compared to see how variation in the wording of the question affected the results. The Free Trade Agreement signed by "Canada and the United States" was about eight points more popular according to the 1988 Canadian Election Study than the one signed by "Brian Mulroney and Ronald Reagan." The difference in responses can provide important insights into what percentage of the population is ambivalent and what considerations might push people in one direction or the other.

Particularly helpful in measuring the general values, culture or attitudes on a particular issue is the use of multi-item scales. These take the responses to several different questions and combine them in an index. For example, one could ask a series of questions about government spending and taxation, and produce an index that would sort respondents along a continuum running from "very supportive" to "very opposed." Such a scale is particularly useful when conducting more complex multivariate analyses. Such scales are also useful because they remind us not to make too much of any one result and help combat the illusion of absolute proportions in the population for and against certain policy directions. These scales can also minimize, but do not eliminate, idiosyncratic question wording effects. Of course, indices cannot be compared across time unless exactly the same questions are used in subsequent surveys.

Conclusion

When properly conducted, polls can be extraordinarily useful tools, but one must first articulate clearly what one wants to know, and then take the necessary time to formulate good questions. If used carelessly, polls can easily become little more than crutches for those who refuse to think creatively or rigorously about tough issues. It is also important to keep in mind that there are many other credible manifestations of public opinion than general population polls. The views of interest groups, of the media or elites are often equally or more relevant when addressing some questions. On issues of specialized knowledge, in particular, it might be more useful to consult credible representatives of groups or the informed public than the general population.

The following 10 relatively simple questions can help anyone assess a particular poll.

1. Have the exact questions in the poll been asked in the past and what were the results?
2. Have similar questions been asked recently and what were the results?
3. What type of poll is it (omnibus, commissioned, academic)?
4. What were the exact dates of polling?
5. Who conducted the poll, for whom, and for what purpose?
6. What were the exact question wordings?
7. What was the order of questions?
8. How were undecideds treated?
9. What was the response rate?
10. Is this really something the public has views about?

Matthew Mendelsohn is Associate Professor of Political Science and Director of the Canadian Public Opinion Archive at Queen's University (mattmen@politics.queensu.ca). **Jason Brent** is a Partner at the I-poll Research Group (jason.brent@ipollresearch.com). They thank Patrick Kennedy for his help with this essay.

Endnote

[1] To calculate the margin of error, one should divide .5 by the square root of the number of respondents (giving you the "standard error"), and then multiply this by 1.96. This gives you the "margin of error" for the 95% confidence interval ("19 times out of 20")—in the case of a sample of 1000, the margin of error would be 3.1%. To find the margin of error for the 99% confidence interval, multiply the standard error by 2.56 (99/100 samples taken from the population will be +/- that number)—in the case of a sample of 1000, the margin of error would be 4.0%. This is the easiest way for the non-specialist to calculate the approximate margin of error.