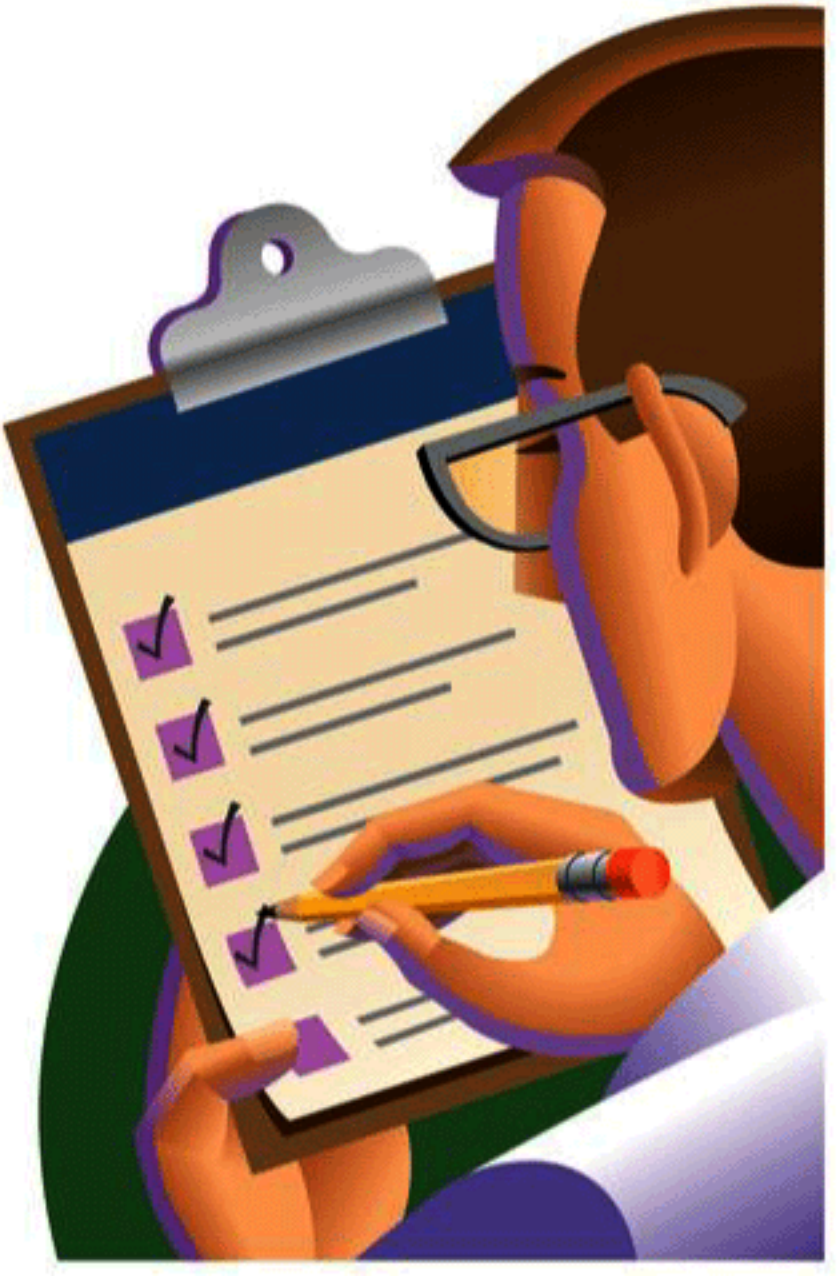
Planning and Using Survey Research Projects

A Guide for Grantees of
The Robert Wood Johnson Foundation



Planning and Using Survey Research Projects:

A Guide for Grantees of The Robert Wood Johnson Foundation

Prepared by Diane Colasanto, Ph.D.,

Princeton Survey Research Associates (PSRA)

Table of Contents

Introduction

The purpose of this guide is to help grantees:

- Make decisions about their survey research projects.
- Know what questions they should ask of the survey experts they hire.
- Recognize the circumstances when it may be acceptable to use a less rigorous data collection method than a full probability survey.
- Understand the consequences of making different kinds of survey design tradeoffs.

It will also help grantees distinguish between the aspects of survey design that are essential, and those that are more flexible.

The recommendations in this guide are based on the author's experience in planning, conducting, and evaluating hundreds of surveys. The recommendations are not intended as a substitute for formal training in survey research methods. Users of this guide will not become survey experts after absorbing the information included below. Instead, they will become informed consumers of survey research, able to make the most of their relationships with the survey research professionals they hire.

Users will also want to consult the American Statistical Association's guides on various subjects in survey research. These guides reflect the experience of a large number of survey practitioners with a variety of skills, edited for the general audience. The guides include such topics as:

- What is a survey?
- How to plan a survey.
- Judging the quality of a survey.

Survey research is an excellent way to measure the knowledge, attitudes, feelings, beliefs, and behavior of many different types of people. It is a powerful

tool for answering the kinds of research questions posed by researchers working on grants provided by The Robert Wood Johnson Foundation (RWJF), such as:

- What do older people know about the health benefits of exercise?
- How many young people smoke cigarettes?
- How effective is advertising in changing people's awareness about health problems?

However, surveys can be expensive and, if not conducted properly, their results can be misleading. Designing and implementing a survey involves many individual technical decisions, each of which can affect the survey's cost and its ability to reliably meet the research objectives.

No survey is perfect. All involve tradeoffs. The need for measurement accuracy and the ability of the sample to represent a larger population must be weighed against time, money and feasibility. Sometimes it is difficult for a person who is not a survey expert to judge these tradeoffs, and even to know whether conducting a survey is the best approach to answering a particular research question.

The first step in the research process is always to have a clear and specific statement of objectives. What questions does the research hope to answer? What is the research trying to accomplish? All later decisions about the technical aspects of the survey's design and about tradeoffs can only be made if the research objectives are firmly in mind.

Grantees planning or purchasing a survey are more likely to get a high-quality product if they budget and plan for the following features:

- Creating a detailed request-for-proposal (see page 30)
- Choosing the best mode of data collection (see page 27)

- Developing a good questionnaire (see page 22)
- Using a “probability sample” (see page 3)
- Choosing an appropriate sample size (see page 7)
- Obtaining a high response rate (see page 16)

At the same time, it is clear that real-world constraints will affect quality. Some general procedures to assess the effects of these constraints and to interpret the results include:

- Ways to compensate for sample design problems (see page 5)
- Methods for over-sampling (see page 11)
- How to achieve an acceptable response rate (see page 16)
- Pre-testing a questionnaire (see page 26)

Probability Sampling Methods

The statistical tests survey researchers use to determine the statistical significance of their results are based on an assumption that the survey data were generated through the use of a "probability sample." This assures that the sample does not contain a *systematic bias* that might skew the survey results away from an accurate representation of the population. The use of probability sampling methods does not ensure that samples will always be perfect representations of the population, since deviations, even large ones, can occur by chance alone. It simply ensures that bias is not built into the design of the sample from the outset. (Bias due to factors other than sampling is discussed later. You may also consult the American Statistical Association guides for further information.)

A probability sample is a sample where every unit in the population has a *known, non-zero, probability* of being included in the sample. If the probabilities of selection are zero, the sample has a problem with the "coverage" of its "frame." If the probabilities of selection are unknown, then the sample is not "representative." Both problems create a strong risk for the sample to have a systematic bias and to lead to erroneous conclusions.

A probability sample is usually achieved by picking units *randomly* from a list of the entire population and giving every unit an *equal* chance of being selected (such as when names are drawn from a hat). The probability does not have to be equal, but it does need to be *known*. Some RWJF-funded surveys sample certain subgroups at different rates, using unequal probabilities of selection. If the requirements of a probability sample are not met, then, technically speaking, one cannot use statistical tests on the data and cannot make any generalizations about the meaning of the results for a broader population than the units actually observed in the survey.

In many real-world situations, strict probability sampling may simply not be feasible because the cost, time and effort involved in doing so are simply too great. How far can researchers deviate from the criteria of probability sampling and still make generalizations from their data? This is an area of some controversy. Researchers and their peers need to judge for themselves whether statistical testing and generalizations are appropriate given the particular sampling compromises they have had to make. The careful researcher will always consider the limitations of his or her own data and will *disclose the details* of how the sampling was accomplished so that others can also judge the reasonableness of the generalizations made from the research.

Researchers who want to generalize their results should strive to accomplish two key goals:
(1) *Complete coverage* of the population so no segment of the population of interest is excluded from the sample.
(2) Making selections at *every* stage of the sampling process through a *random process*, rather than allowing someone's personal judgment (the researcher's, an interviewer's, a respondent's) to determine whether a particular individual is included or excluded from the sample.

Some common ways the principles of probability sampling are disregarded in surveys include:

- People who live in households without a telephone, or who rely completely on cell phone service, are excluded from surveys based on random-digit-dialing or on sampling from telephone numbers listed in telephone books.

- People who live in households with unlisted numbers are excluded from surveys based on sampling from listed telephone numbers.

- People who do not speak English, or who are in poor health, are excluded from most surveys.

- Survey respondents are often selected only from among household members who happen to be at home when the interviewer calls.

- Some surveys expend no effort, or little effort, to locate and convince difficult-to-reach members of the sample to participate in the survey.[1]

- Interviewers in some in-person surveys pick respondents in order to fill quotas for particular types of respondents within each randomly selected geographic area.

- Some surveys recruit participants through advertising, or through networking, or through contacts with clubs and organizations.

There are a number of ways to compensate for many of these sample design problems:

- The researcher can be careful only to generalize results to the part of the population that was covered in the survey, e.g., to the population of people living in telephone households, or to people who speak English.

- The researcher can provide evidence to demonstrate that the sample's shortcomings do not create any systematic biases in the sample, e.g., that the people who were not contacted for the survey are not likely to be meaningfully different from those who were contacted.[2]

[1] Even when a great deal of effort is expended to locate respondents and convince them to participate in the survey, some potential respondents will inevitably choose not to participate. No surveys achieve a 100 percent response rate.

[2] The discussion below on response rates describes some evidence about which aspects of survey non-response create bias in data. There is also evidence to suggest that certain kinds of non-random selection of respondents do not create bias in data. For example, some survey organizations use a method called the

- The researcher can use weights in analyzing the data to compensate for known biases in the sample, giving respondents who are underrepresented a greater weight in analysis than respondents who are overrepresented.[3]

Weights should also always be used in analysis if sample units are selected with unequal probabilities of selection. Remember: the definition of a probability sample only specifies that sampling probabilities are *known and not zero*, not necessarily that they are equal.

In random-digit telephone surveys, people who live in households with more than one telephone number have a higher probability of being included in the sample than those with only one number. People who live alone also have a higher probability of being included in random-digit samples than those who live with other adults, since such surveys generally only interview one person per household. Sometimes, as discussed below in the section on sample size, the researcher will purposefully give some respondents a higher probability of selection than others in order to ensure that there are sufficient numbers of this type of respondent to analyze separately. In all these cases, weights should be used in analysis to compensate for the unequal probabilities of selection and ensure that respondents are weighted to reflect their actual representation in the population of interest.

"youngest male/oldest female" method to select respondents within a randomly-selected household on a systematic, rather than random, basis. The age by gender distribution of these respondents is very close to the Census Bureau's estimate of the distribution of the population. *Caution: this is an area where it is important to rely on experts.*

[3] Weighting can sometimes even be used to compensate for coverage problems, in addition to compensating for non-response problems. For example, respondents who live in households that have not had continuous telephone service throughout an entire year might be given a greater weight in analysis so they also "stand in" for the people who had no telephone service at the time the survey was conducted. There is more on weighting procedures in the section below on response rates. *Again, it is important to rely on experts when dealing with these issues.*

A note on non-probability samples

There is one sampling problem, non-random selection, which cannot be compensated for through analysis or weighting or any other technique. *Research that relies entirely on the self-selection of respondents, or on the researcher's purposeful selection of respondents, can never be used to generalize to a larger population.* Most online polls, call-in polls, and focus groups are based on these types of recruitment techniques where participants select themselves for participation in the research.

It is important to note that this type of research may be useful for certain purposes, despite its severe limitations. Reasonable uses would be to generate ideas for future research, to generate publicity, to make connections with the individuals included in the survey, or to develop qualitative information about a rare population that is extremely difficult to sample using random methods. However, many statistical tests cannot be performed on these data and the results cannot be generalized beyond the individuals actually included in the research.[4]

[4] Quantitative analysis is sometimes useful to perform on data collected through self-selected samples. Percentages, correlations and other quantitative summary measures can be calculated on these data if the purpose of these measures is to describe the characteristics of the sample itself. Tests of the statistical significance of these measures are not appropriate, however. Another exception is laboratory or field experiments on a self-selected population to test hypotheses about causal effects of treatment, but this is a separate purpose than would be served by most surveys.

Sample Size and Over-Sampling

Most surveys are based on interviews with about 1,000 respondents, although many high-quality surveys, including many funded by RWJF, are substantially larger because of the need to cover many geographic areas, many subgroups of interest, and other considerations. However, a sample size of 1,000 has a margin of sampling error of plus or minus 3 percentage points, which is precise enough to suit the purposes of most researchers.[5] A sample of 1,000 cases also allows the researcher to make statistical comparisons among some sub-groups of the population, e.g., comparing the responses of men and women, with a margin of sampling error of about 6 points. When conducting polls that will be reported to the press, this is also a nice, round number that gets the attention of journalists and the public.

In the end, decisions about sample size can only be made by considering the objectives of the research. How precise must estimates generated from the sample be? What statistical comparisons among sub-groups of respondents will be made using these data? Would the researcher be disappointed if the sample was too small to allow a 10-point difference between two groups of respondents, for example, to be considered statistically significant? Are there sub-groups of particular interest that will be analyzed separately from the rest of the sample? How precise must the estimates be for these sub-groups?

Calculating the margin of sampling error

Before deciding what sample size to use, the researcher should calculate what margin of sampling of error will be required for all the important statistical tests to be performed on the data. This is a key technical topic to which the American Statistical Association guides have devoted an entire chapter. In what follows, a

[5] The margin of error reported for surveys is a margin of *sampling* error, not overall survey error. It means: if an infinite number of samples of the same size and design were to be selected, 95 percent of the time the survey estimates would vary from the true population value within the range defined by the margin of error and 5 percent of the time the survey estimates would vary from the population by a larger amount.

beginner should become acquainted with the terminology and then turn to the ASA chapter for further information. The calculation should consider:

- The expected sample size overall and the expected sample size for analytic sub-groups that will be compared.

- The level of confidence that will be used (a 95 percent level of confidence is typical).

- Whether the statistical tests will be one-tailed tests or two-tailed tests (for example, a one-tailed test of the difference between two groups of respondents would hypothesize that one group's response is not *higher* than another group's while a two-tailed test would hypothesize that one group's response is not *different* from another's).

- What the variance will be for the most important survey measures.[6]

- Whether the margin of error needs to be adjusted for the impact of over-sampling or weighting on the statistical efficiency of survey estimates (sometimes called the "design effect" of the sample).

- Whether the margin of error needs to be adjusted for the small size of the total population from which the sample is drawn (this can usually be ignored for populations that contain more than 10,000 units).

Sometimes a researcher will want to conduct separate analyses or comparisons among sub-groups of the population. The margin of sampling error will be much

[6] Previous research can often be used to make these estimates about variance. Most surveys try to make a conservative estimate of variance that can apply to all the measures in a survey. That's why most surveys report a single overall margin of sampling error, rather than a separate margin of error for each survey question. The reported margin of error generally is calculated by making a simple assumption about variance, i.e., that there are two response categories and 50 percent of the respondents choose the first category and 50 percent choose the second.

larger for small sub-groups than for large sub-groups. For example, in the typical random sample of 1,000 adults the results for the sub-group of poor respondents will have a margin of sampling error of plus or minus 10 percentage points, while for higher income respondents the margin will be plus or minus 3 percentage points. In order to conclude that the responses of poor people are different from those with incomes above the poverty line, a difference of at least 11 percentage points between the two groups would be required.

One way to make it easier to conclude that there are statistically significant differences involving relatively small sub-groups of the population is simply to increase the overall size of the sample. For example, in a sample with 3,000 cases overall, the margin of sampling error required to conclude that poor and non-poor people are different reduces to 6 points, from 11. Tripling the sample size, however, is an expensive solution to the problem.

Another, less expensive, option is to *over-sample* the small group. This usually reduces the statistical efficiency of the total sample somewhat, for example to plus or minus 4 points, instead of 3 points, for a total sample size of 1,000.[7] However, this trade-off is usually acceptable if the representation of small, but substantively important, sub-groups of the population can be increased.

An example: deciding about sample size

Here's an example of how decisions about sample size were made for one recent project RWJF funded. AARP was given a grant to conduct a social marketing campaign called *Active for Life*®. The purpose of the campaign was to influence people age 50 and older to exercise at a moderate *level* five days a week, 30 minutes a day. Surveys in two test cities (where the campaign would be implemented) and two control cites (similar to the test cities, but the campaign would *not* be implemented) were to be conducted before, during and after the

[7] If the over-sampled sub-group has much higher variance on the key survey measures than the under-sampled sub-groups, then the statistical efficiency of the sample would be improved, rather than degraded.

campaign to measure the knowledge, beliefs, attitudes, and behavior of people age 50 and older regarding exercise. The success of the campaign would be judged, in part, by whether the survey could conclude that these measures had improved in the test cities over the course of the campaign.

At first, RWJF wanted to conduct 300 interviews in each of the four cities at each of four points in time. The survey contractor for the evaluation estimated that this sample size would make it difficult to conclude the campaign had an impact. For example, an increase of at least 6 percentage points in reports of physical activity for the total sample age 50 and older would have to be observed in a test city (and no change in the control city) in order to conclude that a significant increase in activity had occurred from wave to wave in that city. If RWJF wanted to focus only on the younger adults in the target population, who might be most susceptible to the messages communicated in the campaign, the differences wave to wave would have to be even larger to be statistically significant. The survey contractor felt these required differences set the bar rather high for concluding the campaign had an impact.

Instead the contractor proposed a sample size of 600 per city per wave so it would be easier to conclude that *modest* changes in reported levels of physical activity were the result of the campaign, and not the result of chance variation. At this sample size, only a 4-point increase would be required to conclude that a significant change had occurred. The larger sample size also would allow more flexibility to examine the impact of the campaign on sub-groups of the overall target population, e.g., among people age 50 to 59, among women, and among people with no health problems.

Doubling the sample size for each wave of the survey required some budget trade-offs. RWJF ultimately decided to reduce the number of waves from four to three and eliminate the qualitative research that was included in the original request-for-proposal. RWJF felt the benefits of increasing the sample size for

each wave would be worth more than the information lost by dropping a survey wave and eliminating focus groups. (Eventually, a fourth wave was also funded.)

Methods for over-sampling

Over-sampling is a method for increasing the representation of small, but substantively important, sub-groups of the population over what would be expected from a proportional sample, i.e., a sample with no over-sampling. Over-sampling of a key sub-group can only be accomplished if the target group can be identified in the population *before* the survey is conducted. It is preferable, because it is much less expensive, if the target group can be identified *even before the sample is selected*. Sometimes, however, it is only feasible to have the interviewers identify the members of the target group immediately before conducting the interview. This process of "screening" is the most expensive way to over-sample.

Case A: A population list exists and it contains information about each unit.

In cases where there is a list of the all the members of the population, over-sampling is often easy to do. For example, consider a research project where the population of interest is defined as all the patients in a particular hospital during a specific time period. The patient list to generate the sample would likely include information about the patient's age, duration of stay, and reason for admittance.[8] If the researcher wanted to over-sample young patients, he or she would simply sample patients in different age groups at different rates.

Case B: A population list does not exist, but there is information about groups of units.

[8] Population information such as this can be used to create over-samples of key groups. It can also be used to improve the representativeness of samples that do not include over-sampling. A proportionately-stratified sample uses prior information about the characteristics of the population to divide it into groups, or strata, which are then sampled at *equal* rates to ensure that each group appears in the sample in its *actual* proportion in the population. A disproportionately-stratified sample uses prior information to divide the population into strata which are then sampled at *unequal* rates to ensure that some groups are under-represented and others over-represented in the sample.

Most populations of interest do not exist in the form of a complete list with descriptive information about each unit. For example, most telephone surveys use randomly generated telephone numbers as the basis for the sample, even though people are the units that will form the final sample. In these so-called "RDD" or "random-digit dialed" samples, nothing is known at the sampling stage about the individual people who are the ultimate sample elements. However, information exists about the telephone numbers that could be useful in stratifying the sample, for example telephone numbers can be matched with county of residence.[9]

Companies sell telephone numbers to survey researchers and have compiled extensive information about the characteristics of households served by different telephone exchanges. For example, they can identify telephone exchanges that serve geographic areas with concentrations of African Americans, or Hispanics, or affluent people, or poor people. Samples can be stratified by telephone exchange in order to over-sample exchanges with certain characteristics, and thereby over-sample people with those characteristics.[10]

Case C: There is no advance information about population units

Sometimes a researcher will want to over-sample a particular sub-group of the population, but will neither be able to identify members of the group before the sample is selected, nor be able to use available information, such as geographic location, as a proxy for group membership. In these cases, over-sampling can only be accomplished through *screening*.

Using screening, the researcher contacts a large random sample of the population, interviewing all of the members of the rare sub-group that are identified but only a random subset of the members of other sub-groups of the

[9] All telephone exchanges exist within the boundaries of a particular state and the large majority is assigned within county boundaries. Telephone exchanges usually do not conform very closely to smaller units of geography, including cities and towns, zip codes, and neighborhoods.

[10] Many important characteristics, e.g., age, do not cluster geographically and therefore cannot be over-sampled by stratifying the sample by telephone exchange.

population. This approach saves some expense, because interviews are not conducted with every eligible respondent, but still may be expensive because a large number of potential respondents must be contacted in order to locate enough members of the rare sub-group for the over-sample. Furthermore, some characteristics are difficult to screen on. For example, to locate an over-sample of poor people through screening would require asking a battery of sensitive income questions at the beginning of the interview before the respondent has an opportunity to develop trust in the survey process.[11]

A general note about over-sampling

If you are using a disproportionate sample design as a tool to over-sample important sub-groups of the population of interest, it is very important to remember two points. These are both complex topics that do not lend themselves to an introductory user guide: expert statistical help will be needed.

- All strata, or groups, in the population must be included in the sample design, even if the probability of selection for some strata is very low. Otherwise, results cannot be generalized to the entire population, only to the sampled strata.

- If the difference in selection probabilities across strata or groups is very great, then the overall statistical efficiency of the sample (as represented by the margin of sampling error) may be reduced to an unacceptable extent.

[11] Screening is used in other situations besides the case where a researcher wants to over-sample a particular sub-group. Sometimes the population of interest is defined so that *all* sample members have to be identified through a screening process.

Response Rates

A survey's response rate represents the number of people actually interviewed as a percentage of the total number of people originally sampled and eligible to be included in the survey. Researchers pay attention to the level of the response rate because they believe it is related to the level of "non-response bias." Non-response bias is the extent to which non-respondents and respondents differ on the key behaviors and attitudes being measured in the survey. The relationship between response rate and non-response bias, and how this relationship may vary for different types of surveys, is not yet well understood. Despite this, the response rate is usually the only indicator of sample quality that is reported for a survey. Perhaps this is so because a response rate is easy to calculate. However, to really understand the quality of a sample, it would be important to also examine the amount of non-response bias.

Not only is the response rate usually the only measure of sample quality that is presented in a survey report, unfortunately it usually is also the only measure of *overall survey quality* presented. Many other important aspects of survey quality, such as the reliability and validity of the questionnaire, cannot be summarized in a single quantitative measure that can be compared across surveys. So, the response rate has evolved as a measure of survey quality that is perhaps more important than it is useful.

How to calculate a response rate

Different researchers calculate response rates in different ways. The American Association for Public Opinion Research has undertaken a worthy process to standardize the calculation of response rates so they can be compared across surveys. Their Web site has a document describing the standard calculation and a spreadsheet that makes it easy to perform the calculations.

Regardless of the actual formula used to calculate the rate, an honest response rate always takes into consideration *all* the different ways that the final group of

people who were actually interviewed falls short of the initial sample the researcher targeted to include in the survey. It should take into account the fact that some sampled individuals were not interviewed because they:

- Could not be *contacted* by the researcher. In a telephone survey, these would include such people as those who live in households where the telephone was never answered, or the line was always busy, or an answering machine always picked up the call, or the call was blocked, or the interviewer could not speak the language spoken in the household.

- *Refused* to participate in the survey, perhaps even hanging up on the interviewer in the first few seconds of the call.

- Never *completed* the interview, even though they agreed to participate and started answering the interviewer's questions. Sometimes things come up during the interview and the respondent has to interrupt the survey, and sometimes respondents become annoyed or upset during the interview, or lose interest in participating after answering a few questions.

The rate can be adjusted to compensate for the fact that some of the units in the initial sample turn out not to be individuals at all. For example, most surveys start with a set of randomly generated telephone numbers, not individual people, as the units in the original sample. These samples usually include a lot of non-working telephone numbers that can be excluded from the denominator in calculating a response rate. Business telephone numbers often turn up in random telephone surveys of the general population and these can also be excluded from the calculation.

Other individuals can be excluded from the response rate calculation if they are not part of the survey population under study, i.e., they are *ineligible* to be respondents in a particular survey because they do not meet the screening

criteria that define the population of interest. The response rate is not designed to penalize researchers if their initial sample includes many units that turn out not to be eligible to participate in the survey. However, it is usually difficult for researchers to determine which units are ineligible, since eligibility information is often not available for respondents who are never contacted. In these cases, the researcher has to make his or her own assumptions about how to classify units where eligibility could not be determined. These assumptions will, or course, affect the calculation of the response rate.

What is an acceptable response rate?

Different kinds of surveys achieve different response rates. The U.S. Census Bureau's monthly Current Population Survey usually has a response rate over 90 percent.[12] The most rigorous surveys conducted in the private and non-profit sectors generally achieve response rates in the range of 60 percent to 70 percent. Quick turn-around surveys conducted for media organizations to gauge public response to current events usually have response rates of about 30 percent. Response rates between 40 percent and 50 percent are common for surveys that form the basis of much of what we know about public attitudes and behavior. Are these surveys all equally good?[13]

There is a lot of controversy about the importance of having a high response rate. Some researchers feel that surveys with very low response rates are not reliable at all. Others worry that the techniques used to boost response rates to the 60 percent level and beyond actually introduce more biases into the study's results than they eliminate.

Achieving a high response rate is expensive and time-consuming. High response rates are easier to achieve if respondents have a personal interest in or

[12] See http://www.bls.census.gov/cps/basic/perfmeas/typea.htm.

[13] Large metropolitan areas have lower response rates than small metropolitan areas and rural areas do; this needs to be taken into account when estimating the cost of surveys to be conducted in large cities and in highly urbanized states (e.g., New York).

connection to the survey topic or survey sponsor (as when an association polls its members, or a hospital surveys its patients). But in a typical telephone survey of the general public it would not be unusual for a survey to cost 50 percent to 70 percent more if its goal were a response rate of 60 percent instead of 40 percent. Moreover, it might take three times as long to complete the interviewing on the high response rate survey, compared with the survey with the lower rate.

Is it worth it? Some academic journals only publish articles based on surveys with response rates of at least 50 percent (although such a policy is hardly universal, and a lower response rate may be overlooked in the case of a groundbreaking article by a noted scholar). If your goal is to publish in a journal with such a requirement, you should structure your survey to meet the journal's criterion and spend the extra time and money. Some institutional review boards also may have specific criteria for response rates that you will have to meet.

In other cases, you should:

- *Perform at least a minimal effort to reduce non-response* by making repeated attempts to contact sampled individuals and implementing procedures to persuade reluctant respondents to participate in the survey and complete the interview. For example, telephone surveys with five calls per telephone number and refusal and break-off conversion attempts on the least hostile respondents are neither particularly expensive nor time-consuming.

- *Seek to understand how non-response might affect the survey results* and allocate more resources to minimizing aspects of non-response that will have the greatest impact on the survey's results. You should develop hypotheses about which specific types of non-response are related to the concepts you are studying.

- *Implement procedures to measure the impact of non-response* on the composition of the final sample and consider weighting the data to eliminate the known deviations of the sample from the population. For example, surveys of the general population can be compared against the U.S. Census Bureau's Current Population Survey to determine whether the demographics of the sample match the demographics of the population.[14] If they do not, under-represented respondents can be given a greater weight in analysis so the results of the weighted sample are unbiased.[15] However, it may be difficult to locate reliable population parameters for populations that are not able to be isolated in the Census Bureau's data. In addition, even when data are weighted to compensate for demographic differences between the sample and the population, other sources of bias may exist that are not eliminated by the weighting.

To maximize the response rate, researchers might consider doing the following:[16]

- Make repeated callbacks, up to 20 attempts, at different times of the day and on different days of the week, over a number of weeks. (Some surveys may need more than 20 attempts if they involve a particularly difficult-to-reach population. In general, potential survey contractors should justify their recommendations about callback attempts based on similar studies, rather than on overall company experience or policy.)

- Use special interviewers trained in refusal conversion techniques to follow up with respondents who are reluctant to participate or to complete the interview.

[14] In telephone surveys of the general population, males, young people, African-Americans, Hispanics, and people who did not graduate from high school are typically under-represented.
[15] It is generally best to weight on factors that are relatively fixed and simple to measure, e.g., gender, race, ethnic identity, region of residence, education, number of people in the household. Factors that are difficult to measure, such as income, or that change easily, such as political party identification, are not good candidates for use as weighting factors.
[16] These suggestions apply mostly to telephone surveys, but most can easily be adapted for other types of data collection.

- Send advance notification, where possible, to establish the legitimacy of the survey without revealing too much about its purpose. (Knowing the survey's purpose might create systematic bias in the sample by leading some respondents to decline to participate in the survey.)

- Offer monetary incentives to respondents (although this also has the potential to add bias to a sample, if it is a much greater inducement to poor people than affluent people).

Of course, using a survey questionnaire that is interesting to the respondent, not burdensome to complete, and addresses the respondents' concerns and interests is also a key to a high response rate.

An example: deciding about a target response rate

The survey RWJF funded to evaluate AARP's *Active for Life*® campaign provides a good example of how to allocate survey resources to achieve an appropriate target response rate.

The project started with several constraints — a fixed budget, a fixed schedule (since the "before" survey had to start before the campaign was launched), and a need to conduct a large number of interviews (see the example above on page 10 on deciding about sample size). Within these constraints, it would have been impossible to achieve response rates of 60 percent for each wave of the survey. Either the budget would have to go up, or the sample size would have to go down. The former was not an option, and reducing the sample size would mean that only unreasonably large improvements in respondents' knowledge, attitudes, and behavior would have allowed the researchers to conclude that the campaign had been effective. RWJF did not want to set the bar for the campaign too high, so the survey subcontractor considered what was important in this particular case about achieving a high response rate:

- People who exercise a lot might be difficult for the interviewers to reach at home, so multiple callbacks would be important to include.[17] The contractor decided to conduct up to 15 calls per telephone number to reach potential respondents, because this is the maximum number the contractor thought would be feasible to conduct for the entire sample before the start of the campaign.

- People who have no interest in exercising might become bored or frustrated with a 20-minute interview about exercise and so might break off the interview before completing it. The contractor decided to make an effort to call these respondents back at another time and try to persuade them to finish the interview.

- The researchers had neither a theory nor data to suggest that people who refuse to be interviewed are different regarding exercise than people who are cooperative. Refusals usually take place before the respondent learns anything about what the survey is about, so the contractor figured the decision to participate would be unrelated to feelings about exercise. The contractor decided to do a few refusal conversions, just to be sure about this hunch, but not spend a lot of interviewing resources on trying to get a high cooperation rate.

As it turns out, these decisions were good ones. Analysis showed that the knowledge, attitudes, and behavior of the "converted" refusals were not different from those of the respondents who cooperated without requiring strong persuasion. However, people interviewed after many callbacks were different from those who were contacted on the first few calls to the household. In

[17] The contractor was able to use data from existing surveys to learn that older people interviewed on the first call to a household actually did exercise less often than older people interviewed after many call-backs to the household. It is not always easy to find data to guide these decisions, however, so often you have to decide how to proceed based on common sense.

addition, people who interrupted the interview were different from those who completed the interview on the first attempt. So, the investment in call-backs and in break-off conversions had an impact on the data, while extensive refusal conversions would not have changed the data at all.

The overall response rate for the first two waves of the *Active for Life*® evaluation was 48 percent, because the contractor focused attention on maximizing the two components that mattered, rather than on also maximizing the cooperation rate. The effort to achieve a high contact rate succeeded, with a contact rate of 86 percent. However, as expected, the initial break-off rate on this survey was very high. So, despite having some success in completing partial interviews, the final completion rate, at 90 percent, was lower than it typically is in a general population survey. The cooperation rate was 62 percent.

Another example:
There is other evidence that spending survey resources on increasing the response rate may not affect the data or survey conclusions in a meaningful way:

The Pew Research Center for the People and the Press in Washington, D.C., conducted an experiment in 1997 to try to measure the impact of low response rates on the type of data routinely collected in opinion surveys of the general public. They compared two surveys using identical questionnaires — one completing only five calls per household and minimal refusal conversion attempts and another completing over 20 calls per household and extensive refusal conversion attempts. The first survey achieved a response rate of 36 percent and the second achieved a rate of 61 percent. Despite this large response rate difference, there were very few significant differences between the two surveys on any of the broad range of social and political attitudes measured. In fact, only in the area of racial attitudes did the two surveys differ in important ways.[18]

[18] See http://people-press.org/reports/display.php3?ReportID=94 and *Public Opinion Quarterly* Volume 64 (2000) for two reports of the findings from this experiment. The Pew Research Center is undertaking a similar experiment in 2003-4 to try to replicate this result in a different survey and to explore the impact of

other aspects of survey methodology on survey results. Another example: The University of Michigan researchers who conduct the monthly Index of Consumer Sentiment that is widely reported in newspapers have also found that increasing the response rate does not affect their survey's substantive results (also in *Public Opinion Quarterly* Volume 64). Despite these efforts, the body of knowledge about how response rates affect results is still too slim to be used to develop general principles to guide decision-making in most individual surveys.

Measurement and Testing

Developing a good questionnaire is the most important and difficult aspect of conducting a survey. Although it seems simple, and something that almost anyone could do well, there is considerable art that goes into writing a questionnaire. Seemingly small variations in how survey questions are asked can produce large variations in the answers respondents give. The reader should consult the American Statistical Association Web site chapter on "Designing a Questionnaire" as well as Babbie's excellent book, referenced at the end of this paper.

Training, experience and good judgment are all essential for writing questionnaires that are unbiased, valid and reliable. Surprisingly, expertise in a subject area can be a detriment to writing a good questionnaire. It is often easy for experts to create questions that inadvertently favor their own personal point-of-view on an issue. At the same time it is often difficult for experts to appreciate the limitations of ordinary respondents' understanding of complex issues. However, subject matter experts need to participate in the questionnaire drafting process to ensure that the questions correctly measure the concepts of interest and that the survey objectives are being met.

Rules for writing good questions are even harder to come by than rules for structuring a sample. Furthermore, there's no quantitative measure that neatly summarizes the quality of a questionnaire, as a response rate gives a sense of the quality of the sampling process. Survey questions that yield many "don't know" responses probably are not good measures, so this rate is one indicator of question quality. However, the opposite is not necessarily true; questions with low rate of "don't know" response are not necessarily accurate measures.

In general, people who are cooperative enough to agree to participate in a survey also want to be seen as good, cooperative respondents. Thus, they will be very sensitive to cues from the interviewer, or from the language used in the

questionnaire, that might signal what the "correct" or most valued response is to each question. Much of the work of good questionnaire design and interviewer training involves removing these inadvertent cues and creating a neutral context for soliciting respondents' responses. All respondents must feel comfortable and encouraged to give whatever response is appropriate to their situation.

Measurement error in surveys can also arise because survey questions ask respondents to perform cognitive or memory tasks that are just too difficult.

Here are some guidelines to follow in writing (and evaluating) questionnaires:

General guidelines

- Make sure the questionnaire as a whole gives respondents the opportunity to express their relevant opinions and experiences regarding the survey's key topics. Don't leave questions out that most people will think are essential to understanding an issue.
- Use simple, clear language that all respondents can understand. If words or concepts need to be defined or clarified for respondents, be sure standard language is used for all respondents and that interviewers are not allowed to offer their own explanations to respondents at the spur of the moment.

- Avoid strong or emotion-laden words that might attract or repel respondents to or from particular responses, or might make them feel uncomfortable participating in the survey.

- If a goal of your survey is to measure change over time, be sure to use the same question wording and response categories at each point in time.[19]

[19] It is more difficult to control the question context and other aspects of survey methodology in repeated surveys that measure change, but you should also make every attempt to keep these constant as well.

- Incorporate randomized experiments on question wording and question order into the questionnaire in order to determine the extent to which variation in wording and order might produce different substantive results.

- Use tested questions unless there is a compelling reason to develop new ones. Links to health survey questions such as those on the SHADAC, CDC, and MEPS Web sites can be very helpful in designing questions.

Minimizing respondent burden

- Ask questions respondents can answer without causing them stress, embarrassment, or asking them to perform difficult cognitive or memory tasks. If asking stressful or difficult questions is essential to the objectives of the survey, frame these in such a way to minimize the respondents' burden. For questions on sensitive topics, it may even be appropriate to try to mask the true intention of the question.

- Try to only ask questions of respondents that they would feel are applicable to their situation. Don't burden them by asking irrelevant questions — use question filters to make sure respondents are skipped out of questions that don't apply to them (computer-assisted interviewing makes this simple to do).

- Don't burden respondents by making them sit through a long interview. Twenty minutes of questions on an unfamiliar topic over the telephone will seem burdensome to most people. On the other hand, 40 minutes' worth of questions on a topic dear to the hearts of respondents (asking parents about their children, for example) might not seem burdensome.

- If possible, offer respondents fixed answers to give in response to questions, rather than asking them to formulate their own answers in an open-ended way. There is a lot of variability in respondents' willingness

and ability to do this cognitive work that doesn't necessarily reflect a real difference in opinion or experience. Response to open-ended questions can often add richness and depth to an understanding of the survey's results, but this type of question should be used sparingly.

Order of questions

- Do not assume that respondents are experts in the topic at hand. Establish a context for exploring the survey topics, moving from general questions to specific questions and giving respondents an opportunity to tell you how much they know or care about the topic.

- Be aware that ideas introduced in early survey questions can affect responses to later questions, sometimes causing respondents to narrow their focus to aspects of the issue that have already been raised. Careful attention must always be paid to the "flow" of the questionnaire and the possibility that a particular sequence of questions may focus respondents' thinking in a particular way.

Response categories

- Be sure all possible answers are reflected in the response categories that respondents are offered and that no particular response category seems to have priority over any others.

- Make sure that categories are distinct and non-overlapping. Avoid long lists of categories that would be difficult for respondents to remember.

- Try to present responses in a balanced, symmetrical and fair way, being consistent across response options in the use of examples, arguments, positive and negative words, and other text cues that might cause respondents to choose one response alternative over another. Avoid questions of the "Do you agree that . . ." type because many respondents will tend to answer "yes" automatically in the spirit of cooperation.

- Responses of "I don't know" and "I don't want to answer that question" are acceptable and should always be recorded. Sometimes it may be appropriate to remind respondents that "don't know" is acceptable. In that case, be prepared for lots of respondents to offer that as a response to opinion questions. Many people who might otherwise express an opinion will take the "don't know" option when it is explicitly offered. On the other hand, it is also sometimes appropriate to push respondents a little when they volunteer a "don't know" response to an opinion question, i.e., the interviewer can often ask respondents in which direction their sentiment lies as a way of probing a "don't know" response without seeming rude or offensive.

Pre-testing

By far the most important rule to follow in writing a questionnaire is to TEST IT! First, just read it aloud to hear how it sounds. Then, try it out on people you know who are knowledgeable about the topic and can offer useful feedback on

the completeness, clarity and fairness of the questions. Finally (though you may have to do this more than once), conduct a formal pretest on a small number of actual respondents. Sometimes monitoring as few as 10 or 15 interviews with a random subset of actual respondents will give invaluable feedback on whether the questions make sense, whether they are too burdensome, and whether they give respondents enough opportunity to report what they really think or have experienced.

The reader should also refer to the American Statistical Association brochure on "How to Conduct Pretesting. Finally, when pretesting questions, cognitive testing is increasingly common (e.g., to understand biases in perception of probabilities and the meaning to respondents of modifiers such as "somewhat" and "slightly").

Modes of Data Collection

There are many different ways to collect data, each with their own strengths and weaknesses.

Telephone surveys are most common and are likely to continue to be the dominant mode of data collection for the foreseeable future. It is inexpensive and easy to generate random samples of telephone numbers, and to make repeated callbacks to numbers to try to achieve a high response rate. However, the ability of telephone samples to produce representative samples is increasingly being challenged. Respondents are less and less willing to be interviewed because they see surveys as intrusive and are not sure the personal information they give will be treated confidentially. Also, there are a range of technologies available, from answering machines to call blocking devices, which make it easy for people to screen out calls from survey organizations. The proliferation of cell phones, and the start of a trend toward replacing landlines with cell phones, also challenges the future viability of the telephone survey method.

Another benefit of telephone surveys is that it is easy to standardize and monitor the interviewing process. Computer-assisted telephone interviewing also makes it possible to create questionnaires that are perfectly customized to the situations of individual respondents. Respondents are rarely happy to spend more than 20 minutes answering questions on the telephone, however, making in-depth exploration of a topic difficult. Also, questions requiring visual aids and long, complex questions cannot be asked on telephone surveys.

In-person surveys allow more flexibility in question content, length and use of visual aids, but they are much more expensive to conduct than telephone surveys. Interviewers also give off more inadvertent cues to respondents when they are sitting face-to-face than when they are speaking over the telephone.

So-called “area probability sampling methods” can be used to generate random samples for in-person surveys, but interviews must be clustered geographically in order to keep the interviewing relatively cost-effective. This clustering reduces the statistical efficiency of the sample. Also, making repeated callbacks to households to ensure a high response rate can become extremely expensive in an in-person interview.

Mail surveys are very inexpensive to conduct because no interviewers are involved in the data collection process. Repeated contacts can be made for very little additional cost. But mail surveys can only be used in cases where a complete population list exists with addresses of population members. So, they are inappropriate for most surveys of the general population. They also are less appropriate for samples that are not highly literate.

Mail surveys have the same flexibility of question content that in-person surveys have, but there is no interviewer available to assist respondents who may be having difficulty figuring out how to complete the survey. Mail surveys should probably not be used in circumstances where complex skip patterns, or many open-ended questions, are needed. There is also no assurance that a mail survey will be completed by the intended respondent, or that the initial questions in the survey will be answered before the respondent looks ahead to see all the questions in the survey.

Web surveys have many of the same advantages and disadvantages as mail surveys, but these are heightened for web surveys. They are very inexpensive to conduct, but the sampling problems are formidable. Web surveys are only appropriate for people who are computer literate and have access to the Internet. Web surveys, however, have the additional advantage of being able to be completed very quickly.

Computer programming can ensure more control of how the questionnaire is answered, and by whom, than in a mail survey. At this point, the lack of a simple way to generate random samples of e-mail addresses from a meaningful population frame limits the usefulness of web surveys for surveys of all but very special populations.

How to Evaluate a Survey Research Subcontractor

This guidebook can help you identify many of the pitfalls in conducting survey research and will help you ask good questions about surveys. However, it is not a "how to" guide for conducting survey research. So, unless you have some survey research training and experience, it is a good idea to use a professional survey researcher as a subcontractor.

You may be familiar with various survey organizations through their work that has been published or reported in the media. You can also locate reputable firms through professional associations such as the American Association for Public Opinion Research and the American Marketing Association. Many colleges and universities also have survey research units on their campuses.

The Request for Proposal (RFP)

Your choice of a survey research sub-contractor will be easier if you solicit bids and proposals from a few different vendors *and* if you are very specific in your request for proposal (RFP) about your research objectives, specifications and constraints (especially constraints on budget and time). You need not specify every detail of your project in advance. In fact, it may be more useful to solicit the advice of each firm about how they would structure the research to meet the objectives (but don't go overboard here — vendors are sometimes reluctant to invest a lot of time and energy to provide free survey research design consultation in a proposal).

An RFP is a standardized request that asks vendors to describe their experience and qualifications to conduct your research and to provide a plan, timeline and budget for completing the work. The more specific you can be about your expectations for the proposal and the research, the easier it will be for you to compare proposals and decide among vendors. Be sure you are clear about the final deliverables you expect to receive.

If you do not lay out the research specifications in detail, be aware that you may not be able to make direct comparisons of the cost proposals you receive. A proposal that is very attractive on cost may have used different assumptions about how the research would be conducted than a more expensive proposal. Before you make a decision on cost alone, make sure you ask follow-up questions on the assumptions behind each bid.

Here are examples of some specific questions you might ask in an RFP:

- Is the proposed sample size adequate to meet the objectives of the analysis?

- What is the expected margin of sampling error overall and for key statistical tests?

- Is the proposed length of interview adequate to meet the objectives of the research?

- Are there any survey concepts that will be difficult to measure, and how might these measurement problems be solved?

- What are the procedures for testing the questionnaire?

- How are interviewers recruited, trained, monitored and supervised?

- What are the possibilities for you, the researcher, to monitor interviews yourself?

- If you have a specific response-rate goal, what procedures will the contractor use to achieve the goal? (Be sure these are described in detail: number of attempted contacts with respondents, type of contact, and timing of contacts.)

- If you do not have a specific response-rate goal, approximately what response rate does the contractor expect to achieve with the procedures proposed? (Be sure these are described in detail: see above.)

- Are there any particular types of non-response that the contractor thinks present a problem for the validity of the study? If so, how will these be minimized?

- How will the final response rate be calculated?

- How will the representativeness of the final sample be judged?

- What kind of analysis and reporting will be delivered at the completion of the project to assess the quality of the data and potential sources of bias?

- Will the sample be weighted? If so, how will weighting parameters be derived and how will the weights be calculated?

- What are the subcontractor's procedures for coding and data processing?

- Ask for a detailed timeline of when the different stages of the project will be completed.

- Ask for samples of each vendor's work, especially including reports if you expect the vendor to write a substantive report of findings for your project.

- Ask also for client references so you can learn something about each vendor's responsiveness to client needs, flexibility and fairness in solving the inevitable problems that arise over the course of a project, and attention to detail and deadlines.

Additional Resources

Other Short Guides for Further Reference

A useful brochure prepared by the American Association for Public Opinion Research is "*Best Practices for Survey and Public Opinion Research."* It is available at www.aapor.org.

Another useful publication is the National Council on Public Poll's *"20 Questions a Journalist Should Ask About Poll Results."* It is available at www.ncpp.org.

More Extensive Sources for Further Reading

The Practice of Social Research, by Earl Babbie

"Mail and Internet Surveys: The Tailored Design Method," by Don A. Dillman

"Improving Survey Questions: Design and Evaluation," by Floyd J. Fowler Jr.

About the author:

Diane Colasanto, Ph.D., is a board member and founding partner of Princeton Survey Research Associates (PRSA), a survey design and analysis firm that conducts political and social research for media companies, nonprofit associations, government agencies and corporations.

Prior to PSRA's founding, Colasanto was a senior vice president at The Gallup Organization. She joined Gallup in 1983 as chief methodologist, with responsibility for all aspects of sampling, data quality, and complex statistical analysis. In 1985 she became a vice president, and the following year was named director of the communication and policy research division. In 1987 she was promoted to senior vice president.

Colasanto is a former president of the American Association for Public Opinion Research and has served several terms on the association's executive committee. She is also a trustee of the National Council on Public Polls. She has a Ph.D. in sociology from The University of Michigan.

The author would like to thank Jonathan Best of PSRA, Dr. Elizabeth Martin of the U.S. Census Bureau Laura Leviton, Ph.D., of The Robert Wood Johnson Foundation and several anonymous reviewers for their useful comments on an earlier draft of this guidebook.

The preparation of this guidebook was funded through a grant from The Robert Wood Johnson Foundation.