AAPOR REPORT

EVALUATING SURVEY QUALITY IN TODAY'S COMPLEX ENVIRONMENT

May 12, 2016

Project Team

Reg Baker, Co-Chair, Marketing Research Institute International
Mike Brick, Co-Chair, Westat
Scott Keeter, Co-Chair, Pew Research Center
Paul Biemer, RTI International
Courtney Kennedy, Pew Research Center
Frauke Kreuter, University of Maryland
Andrew Mercer, Pew Research Center
George Terhanian, NPD Group

The Co-Chairs extend their special thanks to Anthony Salvanto of CBS News for his help with this Task Force.

CONTENTS

INTRODUCTION AND SCOPE

The ways in which public opinion, attitudes, and behaviors are formed, expressed, conceptualized, and measured are now more diverse than ever. Today's practitioners and consumers of survey data are exposed to a wide array of methodologies, each with its own challenges--increasing costs, under-coverage, low participation, mixing of modes, uncertainty in the links between theory and practical application, etc. Many users of survey data are skeptical of contemporary survey methods, whether new, innovative approaches or more traditional ones. Experts can and do disagree (See, for example, Gelman 2014; Yeager et al. 2011).

At the same time, the public's appetite for surveys has never been stronger, the results of polls and surveys never more closely watched or more widely used. Innovators are taking advantage of the growth in new technologies, exploring new data sources, and experimenting with new methods. There now are more ways than ever to collect useful data. And while a well designed and carefully executed survey still can deliver valid and reliable results, methods and details matter more than ever.

It is in this context that AAPOR formed a task force to examine the current state of survey methods and provide guidance on the types of information survey practitioners and end users need in order to assess the quality and reliability of survey data. AAPOR has provided detailed treatment of many of these issues in earlier task force reports (Baker et al. 2013; Baker et al. 2010). But the survey research world is changing rapidly and it is critical that periodically we take a broader look at the changes afoot and offer guidance to practitioners and consumers on how to navigate the evolving landscape.

This document is comprised of a series of 17 questions that users of survey data should ask to help them make judgments about the validity of a survey's results regardless of the method used. The answers are not definitive since the underlying issues are still much debated. Rather, they identify the types of design and implementation decisions that can have a significant impact on data outcomes. They are meant to alert the data user to potential sources of bias that might be present. In that sense they form a framework for assessing data quality from virtually any survey, while leaving the ultimate judgment about a survey's usefulness to the data user in the context of the decision to be made or the phenomenon being studied.

EVALUATION FRAMEWORK

This report is organized loosely by the major sources of error specified in the Total Survey Error (TSE) paradigm, focusing on considerations of coverage, sampling, nonresponse, and measurement (see, for example, Groves 1989; Biemer 2010). TSE has proven to be an especially useful tool for survey design and assessment. A substantial literature has developed around its use and, more recently, adaptation to new survey methods. Users of survey data are advised to become familiar with its features and the perspective it brings to the increasingly difficult task of making judgments about the validity of the data from any given survey.

Using the TSE framework we offer what we believe are the most critical questions that survey consumers should ask when evaluating survey results. We recognize that a data user's tolerance for error may vary depending on the use to which the data will be put or the decision it is meant to inform, a perspective sometimes called "fitness for use." Some users may require very precise estimates, while others may be content with directional information that can be used in conjunction with other data sources.

This document is intentionally non-technical to make it accessible to as wide an audience as possible. However, there are occasional uses of technical terms and the reader may find it useful to

refer to AAPOR's *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (AAPOR 2016) for more detail.

THE IMPORTANCE OF TRANSPARENCY

Transparency in all phases of a study is essential if we are to fully assess survey quality. Data users require detailed information about how the survey was designed and executed. The AAPOR Survey Disclosure Checklist (2009) offers a minimum standard for the elements that should be disclosed. A more complete set of disclosure standards is set forth in the AAPOR Code of Professional Ethics and Practices (2015). Any reputable survey organization should be prepared to supply this expanded set of information on request, thereby providing survey users with the essential information needed to assess survey quality.

COVERAGE

One of the most important issues to consider is coverage, that is, the extent the sampling frame (i.e., list of units from which the sample is selected) includes all or mostly all of the units in the survey's defined target population There are several questions to ask to understand the potential bias due to coverage issues:

1. Did most members of the target population that the sample is meant to represent have a chance to be selected? If not, are those who did not have a chance to be selected different in important ways from those who did?

Every survey has some target population or group of people (e.g., the general population, likely voters, members of an association, people likely to buy a product) that it is meant to represent. In an ideal world, all or nearly all of the target population would be identifiable and accessible. In other words, there would be a comprehensive list of all target population members (i.e., a sample frame) available. Some commonly used sample frames are considered to represent (or nearly represent) specific populations. For example, if conducting a survey of U.S. households, the use of the United States Postal Service Delivery Sequence File is generally expected to be a nearly complete and accurate sample frame to represent that population. Other complete (or nearly complete) sample frames might be a full email directory of all employees used for an employee survey or email addresses for everyone who participated in an online webinar for an evaluation survey. If the sample for a study is pulled from the full population of that study, then one can feel confident that there will be minimal coverage bias.

In reality, this is seldom the case and researchers make many different choices to address situations where the sample frame is imperfect or non-existent. For example, while the landline telephone frame provided a high level of coverage of households in the U.S. from the 1980s through the 1990s, the dramatic increase in cellphone-only households over the past decade has resulted in that frame now covering a far smaller proportion of households than in the past. To maximize population coverage in areas where cellphone-only households have become common, a dual frame approach that includes both a landline and a cellphone frame is almost always necessary. Those who have no phone access at all are still excluded, although in the US this is typically a very small percentage of the general population. In order to more accurately represent the population, the study also needs to address the potential for double counting, that is, including a person or household more than once in the survey. Dialing cellphones also is more expensive than dialing landlines.

Sampling theory tells us that bias in survey estimates may occur when some portion of the population is omitted from the initial sample frame, particularly when those excluded differ in important ways from those who are included. From the example above, survey results may be

biased if those living in landline-only households differ in their views and behaviors from those who live in cellphone-only households.

Online research, now the most popular method of survey data collection worldwide (ESOMAR 2015), faces its own challenges. An obvious issue is that people who are not online have no chance of being included in a sample. Further, non-probability online panels do not have a uniform method for obtaining respondents and this makes it difficult to evaluate the sampling process and its outcomes. Understanding (a) how they differ from target populations of interest, such as the general population or all online users, and (b) whether they can represent those populations accurately, requires detailed knowledge about the research subject as well as any adjustments made to account for coverage issues.

Other lists, such as customers who bought a specific product or registered for a loyalty program, students in a class, sufferers of a disease, or members of an organization may be less than complete or otherwise imperfect as well. There may be gaps, duplicates, and even errors.

2. If the sample did not come from a traditional sampling frame, how were potential respondents identified and recruited?

Many studies do not have the benefit of a sample frame containing all or nearly all of the target population from which a sample can be drawn. In such cases the researcher selects a sample through other means. This may involve some form of self-selection, that is, a process where identification and/or recruitment is based on the choice of the individual to be surveyed rather than the researcher. One key difference between this approach and traditional probability sampling is that there may be no true sample frame or list containing all or nearly all members of the target population. Another is that selection is non-random, making it impossible to know the probability or likelihood of any particular individual being included in a study.

Studies using online panels have come under scrutiny because most do not start with a list or frame of known potentially eligible emails, people, or households. Rather, they construct a frame by allowing potential respondents to opt into the panel from which samples for individual studies are drawn. The obvious coverage issue is that people who do not have online access or have online access but have not joined the panel have no chance of being included in the sample. A number of online panels solve the coverage problem by recruiting by phone or mail using classic probability sampling methods. Where sample members do not have online access, the panel provides it or conducts interviews with those individuals using other modes to maximize coverage. These panels generally have fewer members than non-probability panels and are considerably more expensive, which has limited their use. Nonetheless, new probability-based panels are growing in number and popularity as the cost of cross-sectional probability surveys continues to rise.

When working with surveys that use online panels it is important to know whether the panel is non-probability or probability based.

Some sophisticated techniques have been developed in recent years to reduce the potential bias from coverage error when using non-probability panels. They typically rely on large panel designs in conjunction with a series of new sampling and adjustment techniques. The specific techniques can vary widely across organizations (see, for example, YouGov 2016; Survey Monkey 2016).

SAMPLING

Once a sampling frame is determined or created, a sample can be selected. Sometimes the entire population can be surveyed (a census) and sampling questions do not apply. When a sample is selected, it is a step where bias can occur, potentially affecting the validity of the survey results. A recent evaluation of nine online panels, including one probability-based panel, showed that

accuracy can vary depending on the sampling method and/or the sample supplier (Kennedy et al. 2016). This area is one that is only now beginning to be considered more fully.

When reviewing a survey, ask these questions about potential sample red flags:

3. How was the sample selected?

Sampling methods can be divided into two broad groups: (a) probability sampling and (b) non-probability sampling. For probability methods, a probability of selection is assigned (explicitly or implicitly) to every individual on the frame. A sampling process is implemented that selects the sample according to these pre-assigned probabilities. Using Neyman's (1934) theory, researchers can determine the likelihood that statistical findings from the sample are the result of chance (though nonresponse can reduce the accuracy of the findings; more on this later).

For non-probability methods, sampling proceeds without known probabilities. Instead, the survey developer makes the decision as to who is in and who is out (e.g., judgment sampling) or the potential respondents may make the decision (e.g., self-selection). Sometimes the methods are used in combination. The choice is often based on whatever sample can be acquired (e.g., there is no way to identify a full listing of a target population) or whether budgets, timing, or methodological needs require a more convenient sample approach (e.g., it is much less expensive and faster to survey self-selected sample of online panelists than reach out to a random sample of households by telephone). Lacking a probability sample, selection probabilities are unknown and, thus, model assumptions are required to adjust the final survey data before making inferences to a broader population. One common approach is to post-stratify the data to known population totals as is often done with probability samples to adjust for nonresponse. It is hard to know the extent to which this method is effective at reducing bias with non-probability samples.

The success of the statistical modeling approach rests squarely on the assumptions of the model and how well they are satisfied for the data at hand. (See Baker et al. 2013 for a discussion of such methods.) However, the full set of information needed to address these questions with a high degree of confidence is not always available.

Regardless of the method used to develop an online panel or non-probability sample it is common practice to collect background information about panelists that can be used for more efficient sampling or screening for eligibility. This potential advantage may be offset by too-frequent survey participation, which can condition panelists to respond in ways that bias survey results. River sampling is an alternative method where potential respondents are recruited directly from a website into a survey without joining a panel. This avoids the risk of conditioning, but little is known about potential respondents before they are sampled. There are few published studies comparing these two sampling strategies. However, those that have been published have found few differences in respondent characteristics (Bremer 2013; Clark et al. 2015.)

More recently, online sample providers have moved aggressively to the use of sample routers—software that screens potential respondents and assigns them to one of many waiting surveys. (See, for example, Santus et al. 2015). This can improve the efficiency of fieldwork, but there is little published research on the effect of routers on bias. One concern is the degree to which sample composition is affected by other surveys that are active in the router at the same time. One simulation study comparing alternative routing protocols found few effects (Brigham et al 2014). This is an area where more research is needed.

4. What steps were taken as part of the sampling and/or data collection process to ensure that the sample is representative of the target population?

One commonly used technique aimed at ensuring that a sample represents the target population is some combination of quota sampling and *post hoc* adjustment (e.g., weighting).[1] These methods use known characteristics of the target population (such as demographics, geographic location, or other behavioral characteristics) to ensure that the (weighted) distribution of these characteristics in the sample is in the same proportions as in the target population.

A key weakness of these approaches is that they generally do not address other unknown characteristics of the target population (such as attitudes) that may be related both to the topic of the survey and the likelihood of inclusion in the survey. For example, individuals who are distrustful of sharing information may be reluctant to join an online panel or participate in a survey where they are expected to disclose personal details about themselves. And, of course, in virtually every country there is a sometimes substantial portion of the general population that does not use the Internet. Relying only on weighting to known demographics to minimize bias at the sampling or *post hoc* adjustment stage is unlikely to yield a sample in which attitudes toward topics such as online privacy are distributed in the same proportion within the sample as in the general population (Kennedy et al. 2016).

In recent years more sophisticated sampling and adjustment techniques have emerged that use a broad range of characteristics to minimize bias through the use of statistical matching (see, for example, Rivers 2007; Terhanian and Bremer 2012; Wang et al. 2015; Rassler 2002). All aim to match a non-probability sample with similar respondents from a high-quality probability sample in order to ensure that the sample distribution is consistent with the target population. Several recent studies have attempted to empirically evaluate this approach. Buskirk and Dutwin (2015) found matching provided a slight improvement over weighting in reducing error when compared to a reference survey. DiSogra et al. (2015) and Burkey et al. (2015) found that matching failed to remove bias in estimates of vaccination rates among small populations, although neither study compared matching to conventional weighting. Each matching and weighting adjustment requires a set of assumptions that may or may not be satisfied in a particular study.

Judging the effectiveness of statistical matching in the context of an individual survey can be difficult for those without a deep understanding of statistics and the survey topic. Nevertheless, similar techniques have been used to good effect in other sectors to understand, for example, how to estimate causal effects (of, say, smoking cigarettes on length of life) in the absence of randomized controlled experimentation (Rosenbaum and Rubin, 1983; 1984). As Baker et al. (2013) observed, they also have application in survey research with non-probability samples but these are not fully developed at this time.

5. How can I tell if these steps were effective?

The standard approach for assessing potential bias is to compare the characteristics of the respondents to the characteristics of the population or some widely recognized gold standard like a very high response rate survey. In probability sampling, this has been attempted by looking at differences in characteristics of responding and non-responding individuals using data from the

[1] Although widely used in online sampling, quotas and weights are not unique to that method. For example, many telephone surveys, especially in market research, rely on quotas. Weighting, especially demographic weighting, is commonly practiced across all types of surveys and sectors, including traditional probability samples.

sampling frame. The topic is covered in many books and articles and often multiple methods of investigating potential bias are recommended because no one method is very informative.

With non-probability samples care must be taken in these types of analyses, especially when quotas are used in sampling. For example, if the sample was designed to achieve roughly equal numbers of respondents by sex, the comparison of the number of respondents by sex is only a measure of whether the sampling targets were met. A more meaningful step would be to compare the respondent distribution to the population distribution using variables not controlled in sampling. For example, if education or home ownership was not controlled in sampling, then these types of comparisons will inform decisions about who did or did not participate in the study. If the respondents are much more likely to be home owners than is true in the population, then it is likely that other variables (e.g., age, income, and length of living in the same location) will also differ, and estimates such as these may be biased. As noted above, with more complex and sophisticated sampling methods, these comparisons become more difficult. Epidemiologists face many of the same issues in observational studies and have proposed some useful guidelines for reporting (Von Elm et al. 2007).

6. What about sampling error?

Survey estimates based on samples are inherently variable, and no single sample is likely to produce estimates exactly equal to the target population values. Probability sampling has the ability to produce an estimate of sampling error--one component of total survey error due to the sampling process itself. Providing the sampling error, often in terms of a margin of error, alerts users to the fact that the survey estimate is not necessarily the true population value. This is useful even when the sampling error is not the largest source of error in the estimate. Similarly, estimates from non-probability samples may not be equal to the target population value and some measure of precision or accuracy of the estimates should accompany the estimates. The specific models and methods used to create these measures for a non-probability sample should be clearly stated so that data users can evaluate them. The "credibility interval" is an example that meets these disclosure goals (Ipsos 2012; AAPOR 2012). That said, as of this writing, there is no widely accepted measure of sampling error for non-probability samples.

NONRESPONSE

Unit nonresponse occurs when people or households are sampled but from whom no data are collected. This can be due to any number of factors: refusal by the selected respondent; not being home when contact is made; forgetting to complete the survey during the field period; etc. Here are some of the key considerations related to unit nonresponse when evaluating survey quality:

7. What was the response rate (for a probability sample) or the participation rate (for a non-probability sample)?

For many years the response rate has been viewed as an important measure of survey quality. At its base, the response rate reflects the percentage of those initially sampled who actually completed the interview. From a technical perspective, there are several ways in which this rate can be calculated based on the types of sampled units included in the numerator or denominator (see AAPOR Standard Definitions2015).

In recent years, research involving probability samples has shown that the response rate alone is not a very good predictor of nonresponse bias. One reason is that bias is determined at the item level and depends on the relationship between the item and the response pattern. The meta-analysis by Groves and Peytcheva (2008) found the linear relationship was weak. They also found substantial nonresponse bias in many estimates. Their data also show that the item-level biases are

more variable as response rates decline, suggesting that while low response rates do not automatically mean that bias will occur, they increase the risk of nonresponse bias. While the steady decline in response rates over about the last 20 years has led many to question its value it remains an important metric.

Response rates are misleading when used with surveys that rely on non-probability respondent selection (e.g., opt-in panels). This is because in non-probability samples, the denominator for the ratio (i.e., respondents over all sample units) may not be known. Instead ISO 20252: 2008 recommends the term participation rate, which it defines as "the number of respondents who have provided a usable response divided by the total number of initial personal invitations requesting participation" (ISO 2008). AAPOR's *Standard Definitions* adopted this term in 2011. Thus, the participation rate is not equivalent to the probability sample response rate, which itself is not a good predictor of bias. The participation may have some value as a measure of survey quality but it is more often a measure of the capacity (i.e., the number of possible interviews) of an opt-in panel or other source of respondents.

8. How concerned should I be that not everyone who was selected in turn responded?

The potential for bias in estimates due to nonresponse is important in all types of samples. Surveys that have low response may have a high risk of nonresponse bias on some or even all of the questions asked. Weighting adjustments based upon auxiliary data can reduce this risk, but it is often difficult to know how successful the adjustment was. If auxiliary data are available for those who were selected in the sample, then there are greater opportunities to evaluate the bias due to nonresponse. Unfortunately, this is not often the case.

One way to evaluate the potential for nonresponse bias is to attempt to contact and interview a substantial portion (or a sample) of the nonrespondents. This is often a very costly and time-consuming task and often doesn't yield information on the hard-core nonrespondents

There are situations when it is possible to compare respondents and non-respondents without having to re-contact and interview the original non-respondents. For example, a retailer or airline that maintains a loyalty program might decide to survey its members. Assuming that it stores transaction information (e.g., products purchased, flights flown, amount spent), it might be possible to assess whether survey participants differ from non-participants in known ways by analyzing the transaction information of both groups. In many cases, such as list-based voter samples, researchers will have and make use of individual-level information, and could potentially weight the sample to account for the characteristics of those who refused.

On the other hand, public opinion researchers may not have such information available, in which case assessing whether and to what extent survey respondents represent a broader population on variables other than demographics is far more difficult.

The bottom line is that there is good reason to worry about nonresponse bias in survey estimates particularly if response or participation rates are low. The less that is known about the selection process and the outcomes, the greater the concern about the biases. This holds for probability and non-probability samples.

9. How can I tell if nonresponse is a problem, that is, might be leading to bias in the survey results?

Calculating a response rate alone is not enough to determine whether nonresponse will bias your data. To understand why, note that the nonresponse bias for the mean of some characteristic in the survey can be expressed approximately as the nonresponse rate (*nr*) times the difference in the characteristic means of participants and nonparticipants (*d*). To understand the potential for nonresponse bias in a survey it is critical to consider both components. For example, if the level of nonresponse is high and there is a substantial difference in the measured attitudes or behaviors between those who are surveyed and those who are not, then there is a much higher chance of bias in the survey estimates. However, there may also be times when the level of nonresponse is high, yet there are few differences in the attitudes of survey respondents and nonrespondents. In this case, the chance of nonresponse bias may actually be quite low, even though the study response rate is low. A response rate at best, therefore, simply provides a measure of the potential bias, not the actual bias in survey estimates.

Various indicators for risk of nonresponse error have been proposed (see Groves and Peytcheva, 2008; Wagner 2012; Nishimura, Wagner & Elliott 2015). They can be broadly summarized into three methods that enable practitioners to judge whether nonresponse bias is a problem: a) comparison of the survey to external data; b) investigation of internal variation within the data, for example a difference in estimates by the number of contact attempts; and 3) examination of alternative post survey adjusted estimates, where each adjustment is done with different assumptions about the nonresponse process.

All three have strengths and weaknesses and, ideally, multiple approaches are used. What is important to note is that nonresponse bias is specific to a specific estimate, so separate assessments may be needed for different estimates. Thus when deciding on a method, one should consider the available data for each key variable in the data set separately.

Ideally, one has nonresponse bias studies in mind prior to data collection so auxiliary variables that are correlated with both the likelihood of responding and key survey variables can be collected along the way. See Fahimi et al. (2015) for an example.

10. What steps, if any, were taken to adjust for nonresponse?

Numerous methods have been developed to correct for nonresponse in probability samples. Most often, adjustment involves assigning weights to respondents so that cases that are underrepresented relative to the target population contribute more to estimates than cases that are overrepresented. See Brick (2013) for a review of many of the most commonly applied weighting techniques.

Regardless of the specific weighting technique employed, several important considerations should be taken into account when evaluating nonresponse adjustments. The first is the selection of the variables for use in the adjustment. For weighting to eliminate bias, all of the variables that are associated with both response to the survey and to individual questions must be included in the adjustment.

Another important consideration is the data source to which the sample is being compared or matched. Surveys are often weighted so that demographic characteristics match the distributions reported in government surveys or censuses. While the latter surveys often produce high quality benchmarks, it may be the case that successful adjustment depends on use of non-demographic factors that are not measured in government surveys.

Some researchers working with non-probability samples make use of a parallel probability survey that includes non-demographic measures to be used in adjustment (e.g., see Duffy et al. 2005; Terhanian and Bremer 2012). A parallel reference survey allows for the use of variables that would not otherwise be available, although the quality of these measures may be less certain.

Multi-level regression and post-stratification (MRP) is a technique for adjusting non-probability samples. Similar to matching in sample selection, these methods show promise in their ability to reduce selection bias in non-probability surveys. For example, MRP has been shown to produce accurate estimates of presidential voting from a very biased sample drawn from users of the Microsoft Xbox (Wang et al. 2014). On the other hand, Petrin and El-Dash (2015) compared a simple MRP model to conventional post-stratification and found no improvement or only a slight improvement in estimates of political attitudes relative to benchmarks.

The success of these and other adjustments requires the availability of variables that are correlated with both the survey outcome of interest and inclusion in the respondent sample in order to eliminate bias due to nonrandom selection. Earlier comparisons of respondents to probability and non-probability surveys found early adopters of technology to be disproportionately represented in non-probability online samples (DiSogra et al. 2011). More recently Fahimi et al. (2015) identified additional behavioral and attitudinal characteristics that differentiate probability and non-probability respondents that may be useful for adjustment.

As should be clear from the forgoing discussion, researchers face a diverse array of choices with respect to adjustments to both probability and non-probability samples. In the absence of firm guidelines, survey users should ensure that they possess a clear understanding of any proposed methods and consider carefully how they may affect the validity of the research findings. Thus, it is critical that the survey developer identify the technique that was used as well as provide a list of the variables and their source.

Finally, many data providers do not weight data, relying instead on features of the data collection such as quotas to produce a representative sample. If weighting is not applied, the survey provider should provide some evidence that it is not needed.

11. What impact did these adjustments have on the survey results?

The effect of nonresponse adjustments will be different for each item measured in a survey. Comparing an estimate calculated both with and without the use of weights is a simple way to determine the size of their effect. However, this comparison does not necessarily speak to their efficacy at reducing bias. Assessing bias reduction depends instead on the availability of external benchmarks and the background knowledge and expertise of the researcher.

One common effect of weighting is an increase in the variability of estimates. The more extensively a sample is weighted to adjust for selection or nonresponse bias, the more variable the weights. This in turn leads to larger standard errors and reductions in the ability to detect statistically significant results. When using weighted survey data, it is important to use statistical techniques that can account for the effects of weighting on the precision of survey estimates. Most major statistical software packages have routines designed specifically for weighted survey data.

Even with non-probability samples when there is no widely accepted way of computing a margin of sampling error for estimating population proportions and means, it is important to assess the variability of estimates. Such samples are often used to assess the impact of experimental treatments, and the variability of estimates must be accounted for in making such assessments.

As this suggests, correcting nonresponse bias through weighting often comes at a price in terms of increasing the margin of sampling error and decreasing the effective sample size. Any post-survey adjustments of the data should be made with this tradeoff in mind.

MEASUREMENT

Systematic or variable measurement errors can result from a variety of conditions, such as poorly worded questions, the position of a question within a questionnaire (context effect), whether the question is administered by an interviewer or self-administered, the mode in which the survey is administered (i.e., telephone, online, in-person, etc.), and the like. These are some key questions related to measurement that end-users would use to assess survey quality:

12. How was the survey administered (e.g. in person, by telephone, online, multiple modes, etc.)?

The mode used to administer a survey can have an effect on how people answer questions and, in turn, on the estimates. Modes of survey administration vary in their strengths and weaknesses. Self-administered modes, such as web and mail surveys, tend to yield more accurate results when measuring sensitive topics (e.g., voting, drug abuse, sexual behavior). Interviewer-administered modes, such as telephone and in-person, tend to be superior for interviews that are lengthy, feature complicated questions or concepts, or both.

Interviewer-administered modes also tend to be better for surveys focusing on low-literacy populations because they do not rely on respondents reading the questions. Generally, interviewers can increase data quality by keeping respondents engaged and clearing up misunderstandings, but their involvement can undermine data quality when respondents are being asked questions that are potentially embarrassing or related to discernable characteristics of the interviewer (e.g., race or gender).

In addition to measurement, mode of administration can also bear on other sources of error, such as coverage. For example, approximately 2% of Americans do not have a telephone and approximately 10% do not have access to the Internet. These proportions are often significantly larger when doing research outside of the US. Therefore, the mode of administration can carry information about excluded segments of the target population, unless the survey made an effort to cover the excluded individuals (e.g., equipping those without the device).

Survey mode also bears on nonresponse error. For example, young adults are generally less likely to respond by mail and more likely to respond online, relative to older adults. Some surveys leverage these dynamics by offering multiple modes of administration. Offering multiple modes can help reduce nonresponse error, but can also pose a challenge if the mode had an impact on how people responded (as with the sensitive topics mentioned above). Most multi-mode surveys do not adjust for mode effects, but techniques for doing so are available and it is a growing area of research.

13. Were the questions well constructed, clear, and not leading or otherwise biasing?

Survey consumers should be aware that the survey results can be altered by minor changes in wording, the ordering of response options, and the placement of the question within the interview. Questionnaires must be constructed carefully to minimize the amount of error introduced by how the questions are asked. When evaluating a survey question, it may be helpful to consider whether the question:

- Avoids leading respondents to answer a certain way, including statements in support of an issue, cause or organization that may be related to questions
- Asks about just one construct, not two (i.e., double-barreled questions)
- Avoids ambiguity, confusion, and vagueness
- Avoids emotional language

- Uses vocabulary and grammar appropriate to the population be surveyed
- Avoids asking beyond a respondent's capabilities
- Avoids asking about future intentions, although this sometimes is not possible in certain types of surveys*.*

It can also be informative to see where a question was asked in the course of the interview. Preceding questions, typically those immediately preceding, have the potential to influence responses to later questions by drawing the respondent's attention to certain considerations and not others (e.g., Schuman and Presser 1981). Details about which response options were presented versus volunteered and whether or not response options were rotated to mitigate recency or primacy effects can also be useful in evaluating a questionnaire.

To promote transparency on these issues, AAPOR's code, among others, requires researchers to disclose the exact wording and presentation of questions and response options whose results are reported as well as who sponsored the survey and who conducted it. This information can be useful in determining if there was a potential conflict of interest in how the questions were designed.

Finally, questionnaire length also can be important. Long questionnaires can sometimes lead to respondents reducing the cognitive effort they put into answering the survey questions, a phenomenon known as satisficing. In self-administered modes, the mechanics of answering can becoming tiring and further encourage reduced cognitive effort. Unfortunately, there are no hard and fast rules that specify how long is too long, although in general we can assume that self-administered surveys should be shorter than interviewer-administered surveys, telephone surveys shorter than face-to-face surveys, and online surveys (especially if some respondents are likely to complete on a mobile device such as a smartphone) should be the shortest of all.

14. What steps, if any, were taken to ensure that respondents were providing truthful answers to the questions, and were any respondents removed from the final dataset (e.g., identifying speeders, satisficers, multiple completions)?

Over the last decade or so, there have been widely expressed concerns about the in-survey behavior of respondents drawn from online opt-in panels, in particular. The emergence of "professional respondents" who complete large numbers of surveys, "speeders" who complete surveys too quickly, and evidence of random selection or overuse of non-substantive answers have led to the development of techniques to identify and remove data records with these characteristics. In fact, many survey buyers, industry codes of practice, and both ISO 20252 and 26362 require it.

More recently, research has begun to suggest that due to the combination of random responding and generally low incidence of these behaviors in online surveys, the impact of these respondents and behaviors on survey estimates is ignorable (see, for example, Hillygus, Jackson, and Young 2014; Greszki, Meyer, and Schoen 2015). Based on their analysis of a very large and rich dataset collected by the Advertising Research Foundation, Thomas and Barlas (2014) reported similar findings while also showing that these behaviors are often a function of poor questionnaire design. They also showed that removing the data records from the dataset might introduce bias by reducing diversity and making the sample less like the target population.

Researchers working in more traditional offline modes (e.g., telephone, mail) have similar concerns with respect to satisficing, which they generally have classified as a problem to be solved by questionnaire design. Removal of data records in these modes is rare.

OTHER FACTORS:

A number of additional factors can impact survey data quality. Three of the more important involve the length of time the survey was fielded, the use of incentives, and the reputation of the organization conducting the survey:

15. How long was the survey in the field and how much effort was put to ensuring a good response?

As a general rule, the longer a survey is in the field the greater the opportunity to achieve a higher response and a more representative sample, as early responders can sometimes be different from later responders. Longer field periods allow for prompting and follow-up as well as targeting of under-performing demographic groups or subsamples.

That said, there are no hard and fast rules about how long a survey should be in the field. Some methods are inherently faster than others. For example, there is almost no practical limit to the size of sample that can be fielded online in a single day. Surveys that rely on interviewers, on the other hand, are constrained by the number of interviewing hours they can put to the task. Mail surveys often take longer than other methods because of the delays imposed by delivering materials, communicating with respondents, and having completed surveys returned.

There are exceptions to the rule that a longer field period is better and one of the most important involves survey topic. For example, many public opinion surveys deal with volatile topics about which attitudes can change in a matter of days. Political polls are a good example, and typically interviewing within a very tight time period is necessary.

16. What incentives, if any, were respondents offered to encourage participation?

There is a long list of reasons someone might or might not choose to respond to any survey and a host of external factors that may influence them positively or negatively toward a response decision. Many of the issues pertaining to in-person and telephone surveys apply to online surveys as well. Incentives of some form are often used in surveys of all types. These can range from cash to prizes, sweepstakes, donations to a charity, points that accrue over time, etc. Both direct and indirect incentives can increase response rates and improve data quality. However, they can also reduce data quality and increase nonresponse bias depending on a number of factors (e.g., see Göritz 2004; Singer and Ye 2013; Singer and Couper 2008).

When evaluating survey results, it is important to consider the potential impact of employing incentives in a survey. For example, monetary and other direct incentives could bias the survey response toward lower income groups. Further, incentives could affect how respondents respond to individual questions. The promise of an incentive could induce some respondents to rush through the survey only to collect the reward, thus leading to the collection of thoughtless responses. These issues should be considered as risks for all surveys that employ incentives, not only those conducted online.

The absence of incentives – regardless of the type of survey – may introduce other biases and provides no assurance of data quality. In an era of declining response rates, incentives can increase panel retention, improve data quality, and produce higher response rates for groups that otherwise would have been underrepresented. Incentives can also mitigate a bias toward groups with an especially high interest in a subject or increase participation for groups who have little time for surveys. Disentangling all of the complex effects of incentive use is quite difficult and most data users do not have the data to allow such assessments. Nevertheless, awareness of the benefits and risks of incentives provides some defense from being misled about their use.

17. What is the record of accomplishment of the organization that conducted the survey?

Finally, users of survey results should consider the organization's record of accomplishment, its use of best practices, the contributions it has made to the field, its transparency of methods and its record of past accuracy.

Organizations with long and successful records can and should inspire confidence. Equally important is the degree to which their methodologies reflect the most current thinking about the challenges faced by contemporary survey research (e.g., potential issues in non-coverage or recent changes in the population of interest). Newer organizations and innovations should always be welcomed and not avoided simply because they are new. In all cases, full transparency about their methods and assumptions is essential.

For example, often a track record becomes synonymous with how well an organization has performed in forecasting election results. This is understandable, and from the earliest days of public polls has been widely accepted—even welcomed—largely because the survey researcher ultimately sees the true population parameter against which the survey estimate can be compared.

This is not the only criterion against which to evaluate an organization's track record. Studies should be considered for their consistency across a range of measurements. For example, if a study estimates the winner of an election but the demographics of likely voters differ widely from those reported in other studies, then a closer look is warranted.

Organizations that inspire confidence should have a track record of using widely-accepted best practices; of making or seeking improvements when accuracy has fallen short; of cooperating with requests for details about their methodologies; and showing a willingness to contribute to the larger understanding of market, opinion, and social research by contributing to the discipline's broader understanding of survey methods. Organizations that have chosen to join AAPOR's Transparency Initiative have signaled their commitment to many of these principles.

REFERENCES

AAPOR (2009). Survey Disclosure Checklist. Retrieved on August 22, 2015 from http://www.aapor.org/AAPORKentico/Standards-Ethics/AAPOR-Code-of-Ethics/Survey-Disclosure-Checklist.aspx.

_____(2012). AAPOR Statement: Understanding a "credibility interval" and how it differs from the "margin of sampling error" in a public opinion poll. Retrieved on February 5, 2016 from https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/DetailedAAPORstatementoncredibilityintervals.pdf.

___ (2015) AAPOR Code of Professional Ethics and Practices. Retrieved on August 23, 2015 from http://www.aapor.org/AAPORKentico/Standards-Ethics/AAPOR-Code-of-Ethics.aspx.

___(2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Retrieved May 9, 2016 from http://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf.

Baker, R., Blumberg, S.J., Brick, M.J., Couper, M.P., Courtright, M., Dennis, J.M., Dillman, D., Frankel, M.R., Garland, P., Groves, R.M, Kennedy, C., Krosnick, J., Lavrakas, P.J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R.K., Zahs, Z. (2010). AAPOR Report on Online Panels. *Public Opinion Quarterly* 74(4):711–81.

Baker, R, Brick, M.J., Bates, N. A., Battaglia, M., Couper, M.C., Dever, J.A., Gile, K.J., & Tourangeau, R. (2013). Report of the AAPOR Task Force on Non-Probability Sampling. Retrieved on September 7, 2015 from

https://www.aapor.org/AAPORKentico/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf

Biemer, P.P. (2010). Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly,* 74, 5: 817-848.

Bremer, J. (2013). "Research Quality: The Interaction of Sampling and Weighting in Producing a Representative Sample Online: An Excerpt from the ARF's 'Foundations of Quality 2' Initiative." *Journal of Advertising Research* 53(4):363–71.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29, 329-353.

Brigham, Nancy, Michael Fallig, and Chuck Miller. 2014. "The Impact of Survey Routers On Sampling and Surveys: Unraveling the Mysteries of Survey-Router Design and Deployment." *Journal of Advertising Research* 54(4):381–88.

Burkey, A., DiSogra, C., Greby, S., Srinath, K.P., Black, C., Sokolowski, J., Ding, H., Ball, S., Donahue, S. (2015). Matching an Internet Panel Sample of Pregnant Women to a Probability Sample. *Presentation at the 2015 Conference of the American Association for Public Opinion Research*.

Buskirk, T. D., & Dutwin, D. J. (2015). Selected of Self-Selected? Part 2: Exploring Non-Probability and Probability Samples from Response Propensities to Participant Profiles to Outcome Distributions. *Presentation at the 2015 Conference of the American Association for Public Opinion Research*.

Clark, J., Young, C., & Petrin, R. (2015). Meta-Analysis of Online Panel and Non-Panel Sampling: Electoral and Non-Electoral Behavior Metrics. *Presentation at the 2015 Conference of the American Association for Public Opinion Research*.

Disogra, C., Curtiss C., Elisa C., and Dennis, J. (2011). Calibrating Non-Probability Internet Samples with Probability Samples Using Early Adopter Characteristics. Pp. 4501–15 in *JSM Proceedings (Survey Research Methods Section)*. Alexandria, VA: American Statistical Association.

DiSogra, C., Greby, S., Srinath, K. P., Burkey, A., Black, C., Sokolowski, J., Yue, X. Ball, S. Donahue, S. (2015). Matching an Internet Panel Sample of Health Care Personnel to a Probability Sample. *Presentation at the 2015 Conference of the American Association for Public Opinion Research*.

Duffy, B., Smith, K., Terhanian, G., & Bremer, J. (2005). Comparing Data from Online and Face-to-Face Surveys. *International Journal of Market Research*, 47, 6: 615–39.

Fahimi, M., Barlas, F., Thomas, R., & Buttermore, N. (2015). Scientific Surveys Based on Incomplete Sampling Frames and High Rates of Nonresponse. *Survey Practice, 8(6)* Retrieved from http://www.surveypractice.org/index.php/SurveyPractice/article/view/324

Gelman, A (2014). When should we trust polls from non-probability samples?" *The Washington Post*. April 14. Retrieved on January 15, 2016 from https://www.washingtonpost.com/news/monkey-cage/wp/2014/04/11/when-should-we-trust-polls-from-non-probability-samples/.

Göritz, A.S. (2014). The impact of material incentives on response quantity, response quality, sample composition, survey outcome, and cost in online access panels, *International Journal of Market Research*, 46, 3:327-345.

Greszki, R., Meyer, M., & Schoen, H. (2015). Exploring the Effects of Removing 'Too Fast' Responses and Respondents from Web Surveys. *Public Opinion Quarterly,* 79, 2: 471-503.

Groves, R. M. (1989). *Survey Errors and Survey Costs.* New York: John Wiley & Sons.

Groves, R.M., and Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias a meta-analysis. *Public Opinion Quarterly,* 72, 2: 167-189.

Hillygus, S.D., Jackson, N., & Young, M. (2014). Professional respondents in non-probability online panels. in Calegaro, M., Baker, R., Bethlehem, J., Göritz, A.S., Krosnick, J.A., & Lavrakas, P. J. eds. *Online Panel Research: A Data Quality Perspective*. UK: John Wiley & Sons.

Inside Research (2014) Special Report: Worldwide Online Spend Growth Accelerates. *Inside Research,* 25, 3: 4–5.

Ipsos (2012). Credibility Intervals for Online Polling. Retrieved on February 5, 2016 from https://ipsos-na.com/dl/pdf/research/public-affairs/IpsosPA_CredibilityIntervals.pdf.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., and Gimenez, A. (2016). Evaluating online non-probability surveys. Pew Research Center. Retrieved on May 9, 2016 from http://www.pewresearch.org/2016/05/02/evaluating-online-non-probability-surveys/

Little, R. J., & Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology*, 31, 2: 161-168.

Neyman, J. (1934). On Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection (with discussion). Journal of the Royal Statistical Society, 97: 558-625.

Nishimura, R., Wagner, J., & Elliott, M. (2015). Alternative Indicators for the Risk of Non-response Bias: A Simulation Study. *International Statistical Review*, 84, 1: 43-62

Petrin, R. & El-Dash, Neale (2015). Reaching Wider, Going Deeper: Incorporating Sample Source Variation And Other Considerations Into MRP Adjustments Of Polling Estimates From Blended River Samples. *Presentation at the 2015 Conference of the American Association for Public Opinion Research*.

Pew Research Center. (2015). Coverage Error in Internet Surveys. Retrieved September 25, 2015 from http://www.pewresearch.org/files/2015/09/2015-09-22_coverage-error-in-internet-surveys.pdf

Rassler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches.* New York: Springer.

Rivers, D. (2007). Sampling for Web Surveys. White paper prepared from presentation given at the 2007 Joint Statistical Meetings, Salt Lake City, Utah, July-August. Retrieved on January 16, 2016 from https://s3.amazonaws.com/yg-public/Scientific/Sample+Matching_JSM.pdf

Rosenbaum, P.R., and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70, 1: 41-55.

Rosenbaum, P.R., and Rubin, D.B. (1984). Reducing Bias in Observational Studies Using Subclassification on The Propensity Score. *Journal of the American Statistical Association*, 79, 387: 516-524.

Santus, D., Kwok, P.K., and Kelly, F. (2015). Should Sampling Be Left To Chance? Controlling For Non-Quota Variables In A Sample. Retrieved on October 18, 2015 from http://www.lightspeedgmi.com/wp-content/uploads/2015/03/Casro_Paper_controlling_variables.pdf.

Schuman, H. and Presser, S. (1981). Questions and Answers in Attitude Surveys: Experiments in Question Form, Wording, and Context. New York: Academic Press.

Singer, Eleanor, and Mick P. Couper. 2008. "Do Incentives exert undue influence on survey participation? experimental evidence." *Journal of Empirical Research on Human Research Ethics*, 3(3): 49-56

Singer, Eleanor, and C. Ye. 2013. "The Use and Effects of Incentives in Surveys." *Annals of the American Academy of Political and Social Science*, 645(1): 112-141.

Survey Monkey (2016). "NBC News / Survey Monkey Weekly Election Tracking Poll." Available at: http://www.scribd.com/doc/294643292/NBC-News-SurveyMonkey-Weekly-Election-Tracking-Poll

Taylor, H., & Terhanian, G. (1999). Heady Days are Here Again: Online Polling is Rapidly Coming of Age. *Public Perspective*, 10, 4: 20-23.

Terhanian, G. and Bremer, J. (2012). A Smarter Way to Select Respondents for Surveys? *International Journal of Market Research*, 54, 6: 751–80.

Terhanian, G., Smith, R., Bremer, J. & Thomas, R.K. (2001). Exploiting Analytical Advances: Minimizing the Biases Associated with Non-random Samples of Internet Users. Proceedings of the ESOMAR/ARF Worldwide Audience Measurement Conference.

Thomas, R.K. & Barlas, F.M. (2014). Respondents Playing Fast and Loose?: Antecedents and Consequences of Respondent Speed of Completion. Retrieved on August 24, 2015 from http://dc-aapor.org/2014%20conference%20slides/ThomasBarlas.pdf.

Von Elm, E, Altman, D.G., Egger, M., Pocock, S.J., Gøtzsche, P.C., Vandenbroucke, J.P., & Strobe Initiative (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. *Preventive Medicine* 45, 4: 247-251.

Wagner, J. (2012). A comparison of alternative indicators for the risk of nonresponse bias. *Public Opinion Quarterly*, *76*, 555-575.

Wang, W., Rothschild, D., Gopel, S., and Gelman, A. (2015). Forecasting Elections with Non-Representative Polls. *International Journal of Forecasting*, 31, 3; 980–991

Yeager, D.S., J.A. Krosnick, L. Chang, H.S. Javitz, M.S. Levendusky, A. Simpser, and R. Wang. (2011) Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples. *Public Opinion Quarterly* 75:709–47.

YouGov (2016). "The Methodology of the 2016 YouGov/CBS News Battleground Tracker." Available at: https://today.yougov.com/news/2015/09/13/methodology-2016-cbs-news-battleground-tracker/