Public Opinion Quarterly, pp. 872–888

THE FUTURE OF SURVEY SAMPLING

J. MICHAEL BRICK*

Abstract The twentieth century saw a dramatic change in the way information was generated as probability sampling replaced full enumeration. We examine some key events of the past and issues being addressed today to gain perspective on the requirements that might steer changes in the direction of survey sampling in the future. The potential for online Web panels and other methods that do not use probability sampling are considered. Finally, we conclude with some thoughts on how the future of survey sampling might be shaped by unpredictable factors.

Contemplating the potential directions that survey sampling may take in the future requires considering the current state of affairs and how we arrived at this point. Despite occasional issues and controversies, it seems fair to say that survey sampling, and probability sampling in particular, has helped transform our view of society and the issues it faces. Perhaps even more importantly, sampling has changed how we think about obtaining information from a large population. We have seen the transition from full enumeration as the only acceptable method of learning about a population that was the standard in the late nineteenth century to "partial investigations" and finally to a full theory of sampling. Kruskal and Mosteller (1980) describe this transformation.

This change has made it possible to study a wide range of topics about our society in a cost-effective and timely manner. The findings from surveys are generally accepted in popular, legal, and technical areas, although there are still debates about methods of implementation. Election surveys, both those in the run-up to the election and exit polls, demonstrate both the general acceptance and concerns about methods. Election surveys are very public and their accuracy is easily judged; the results of these surveys are highly reported and the

J. MICHAEL BRICK is a Vice President, Associate Director, and senior statistician at Westat, Rockville, MD, USA, and a research professor of the Joint Program in Survey Methodology at the University of Maryland, College Park, MD, USA. He would like to thank the editors for the invitation to prepare this article and for their suggestions and comments. He would also like to thank Pat Dean Brick, Jill Montaquila, and Bruce Allen of Westat for their constructive comments on an earlier draft and their sharing of thoughts on new directions for survey research. *Address correspondence to J. Michael Brick, Westat, 1600 Research Boulevard, Rockville, MD 20850, USA; e-mail: MikeBrick@westat.com.

doi: 10.1093/poq/nfr045

public views them as informative. Despite this, methods of implementing election surveys are often debated. For example, as cell phones became more prevalent, the potential bias in the estimates from pre-election telephone surveys that ignored cell phones was widely debated (Keeter 2006). Even though specific surveys or methodologies may be criticized, the basic principles of sampling are rarely questioned. In retrospect, the broad acceptance of survey sampling theory as a means of making inferences about large populations is a very profound and exciting development.

We begin by defining what we include when we talk about survey *sampling* and the *future*. The three key components of survey sampling are sample selection, data collection, and estimation. The integration of these components is one of the features that distinguishes survey sampling from other areas of statistical research. Other authors in this anniversary issue discuss data collection topics, specifically mode, questionnaire design, and surveys for some important applications. As a result, we try to keep the main focus in this article on the bookends of this survey sampling triad. But the three components are so intertwined that any discussion of the future must include data collection and data collection costs.

For the purposes of this article, we consider survey sampling to be the methods for identifying a set of observations from a population and making inferences about the population from those observations. This definition includes important topics such as coverage of the population, development and construction of the sample, sample selection methods, handling missing data, weighting or analysis of the observations to make estimates of population characteristics, and production of estimates of precision from the observed sample. We concentrate on methods for surveying households and individuals rather than establishment surveys.

The other key concept is the future. The near-term future, say the next 10 years or so, is easier to consider because within this time frame we might not expect drastic changes from the current status. The longer-term future is in many ways more interesting. In the long range, it is reasonable to think that paradigm shifts in survey sampling are feasible, even though such changes may not be likely given that none have occurred in the last 75 years. Longer-term speculation may also be safer because fewer researchers will recall if our predictions do not resemble the course of events. We discuss both time frames, but most of our speculations address the longer-term future.

In the next section, we begin by reviewing some of the past and current directions in sampling. The major changes following Neyman (1934) that saw the overthrow of full enumeration and other methods of partial investigation and the establishment of a paradigm for survey sampling are especially relevant. Other changes in surveying such as the move to telephone sampling in the middle of the twentieth century are also instructive. The context for change is also reviewed. In particular, the requirements for timely and cost-efficient estimates were stimulants for changes that interacted with sampling methods, and cost pressures are even more relevant today. Online panels and other methods of

sampling that are not probability samples are considered, and the conditions needed for these methods to provide acceptable alternatives to probability sampling are assessed. Finally, we conclude with some thoughts on how the future might be shaped by factors that none of us can predict.

Background

Twenty-five years ago, Frankel and Frankel (1987) reviewed the history of survey sampling in the 50th-anniversary issue of *Public Opinion Quarterly*. Their article is still very relevant, and today's readers who are interested in broad outlines of the history of sampling will enjoy their review.

Frankel and Frankel (1987) describe two phases of development: The first phase was the establishment of the basic methods of probability sampling following the paradigm-changing article by Neyman (1934). The second phase was characterized by innovations of the basic methods and extensions to accommodate new technologies like telephone sampling and computerization. This dichotomy still serves us well today, as many of the developments in sampling since 1987 have continued the second phase of innovations and technology. For example, registration-based sampling for election polling (Green and Gerber 2006), and sampling persons on cell phones (AAPOR 2010), are innovative extensions of standard sampling methods. The Web is another technological innovation that has become an important mode for conducting surveys and has been the subject of intensive research. As Couper (2000) observes, no generally accepted method of sampling from the Web has been established at this time—and this remains true a decade later. This sampling problem has spawned a host of different possible solutions (Baker et al. 2010), some of which will be discussed later.

Probability sampling as developed by Neyman (1934) and almost all the survey sampling texts that followed have a very specific framework for making inferences. It assumes that a frame of all units in the population can be constructed, that every unit in this frame has a positive probability of being selected into the sample, and that the probability can be computed for each sampled unit. Once sampled, the characteristics of all the units can be measured accurately. Estimates are produced by using a reference distribution of all possible samples that could have been selected using the same procedures. Many procedures such as multi-stage sampling and differential probability sampling methods satisfy this general framework. Nonresponse, incomplete coverage of the population, and measurement errors are examples of practical issues that violate the pure assumptions of probability sampling.

Other sampling methods have been proposed recently, and some of these do not fall squarely within the probability sampling framework. One example is respondent-driven sampling (Heckathorn 1997), which has been used primarily for surveying rare or hard-to-reach populations. This procedure is a network or snowball sampling process that asks members of the rare group to refer others they

know who are eligible for the survey. A mathematical model is used to analyze the data. Volz and Heckathorn (2008) show that the observed sample is equivalent to a probability sample under specified conditions, but those conditions are rarely satisfied in practice. Two other relatively new sampling methods that use networks but fit more closely within traditional probability sampling methods are adaptive sampling (Thompson and Seber 1996) and indirect sampling (Lavallée 2007). Adaptive designs are typically used for estimating the abundance of rare, clustered populations in the natural sciences and have not been applied as frequently in human populations. For example, Thompson, Ramsey, and Seber (1992) describe how adaptive sampling could be used in trawl surveys of shrimp because shrimp have clustering or schooling tendencies but it is not possible to predict in advance where the shrimp will be concentrated. Using estimates of density from a first-stage sample, the second-stage sampling areas are sampled at different rates to capture more shrimp. Indirect sampling has been used in surveys of households and individuals. Indirect sampling begins with a probability sample of persons or units that are not the target population. It then links the target population to the sampled units so that the probabilities of selection of the target population can be computed. Kalton and Brick (1995) apply this approach in household longitudinal surveys to assign weights to people who move into a sampled household after the first wave. We will return to some of these topics in our speculation about the future of sampling.

Frankel and Frankel (1987) did not discuss the foundational debate on survey sampling inference that roiled the statistical community beginning around the 1960s. As with any newly developing scientific area, questions about the appropriateness of the methods used to make inferences from a sample to a target population were raised and widely debated within the statistical community. Godambe (1966) and Royall (1970) are examples of criticisms of the idea of making inferences about a population based on the probability of selection of sampling units (the design-based approach). They and other researchers favored the use of more typical modeling assumptions used in other areas of statistics (model-based and Bayesian approaches) for inference. Holt and Smith (1979) provide a thoughtful piece that expands further on some of these issues. They highlight a concern that design-based sampling may not be consistent with the conditionality principle, a statistical idea that essentially states that once a sample is observed inferences should be conditioned on that sample rather than on other samples that might have been selected. Hansen, Madow, and Tepping (1983) give a defense of the utility of the design-based approach, and show that minor errors in the assumptions in a model-based approach may have important effects on the bias of the estimates when a large sample is selected. While the foundations issues have not been fully resolved, the design-based approach has remained the dominant one in survey sampling.

Even though most survey inferences rely on the design-based method, model-based methods have become more routinely used in areas such as small area estimation. When the sample size for a probability sample is insufficient to make reliable direct estimates at a low level of geography, small area models are

developed to provide those estimates (Rao 2003). A probability sample is not required in this process since the models are the basis for inference rather than the probabilities of sampling the units. Nevertheless, a probability sample is generally the starting point for small area estimation.

Model-based methods are also needed to account for nonsampling errors in probability samples since the design-based theory assumes perfect execution of the sample design and data collection. One example is the handling of missing data from sampled units (Little and Rubin 2002), where the missingness may be due to nonresponse or undercoverage. The design-based and model-based approaches are combined in this situation; a probability sample is selected and design-based methods of weighting are used to project to the population, but model-based methods are used to account for coverage errors, unit nonresponse, and item nonresponse.

Despite the rapprochement of model- and design-based methods, foundational issues are still being debated today in a different guise. The current discussion is driven by practice rather than by statistical theory, with online panels attracting the most attention. An approach that has been explored with online surveys is the use of alternative estimation methods. The hope is that alternative estimation methods might make it possible to make inferences to a general population even when the mechanism for sampling is not well defined. This idea is clearly a model-based approach even if it is not always stated as such.

Some researchers have opined that the era of probability sampling is over because of the ever increasing costs of data collection combined with losses due to noncoverage and nonresponse. A common refrain is that a probability sample with a low response rate or coverage rate is no "better" than a nonprobability or volunteer sample. The concern goes to the foundations of survey sampling and must be addressed for the future of sampling—whether that future is a continuation of the reign of probability sampling or some other procedure for data collection and inference. Our quest for thinking about how these issues might be resolved in the future begins by looking back.

Statistical Theory, Cost, and Sampling

In reviewing the early history of survey sampling, the efforts of Kaier (1895–96), Bowley (1926), and Neyman (1934) are invariably described as keys to the change from a dependence on full enumeration to sampling (Kruskal and Mosteller 1980; Frankel and Frankel 1987; Bellhouse 2000). This change was profound. For example, Duncan and Shelton (1978) describe the effect of the change in U.S. government data collection activities that occurred as a result of the efforts of these pioneers from about 1900 to 1940 as a revolution. It clearly was a major paradigm shift.

One of the interesting aspects of this history is that the revolution in sampling was driven by the demands of government and official statistics. Kaier (1895–96) was the central character, and the International Statistical Institute (ISI) served as

the primary forum for discussion of the virtues of representative samples. Academics also played a role. Bowley (1926) developed seminal ideas about randomization, and Neyman (1934) provided the statistical coup d'état. However, the timing of events is instructive. Kruskal and Mosteller (1980) observe that "by the time of the 1925 ISI meeting in Rome and after Kaier's death, representative sampling had found an accepted, honorable place among statisticians." This acceptance of sampling happened well before the statistical contributions of Bowley and Neyman.

It also appears that the motivation for the adoption of survey sampling was not simply cost-effectiveness, although cost was important. In the histories of the early developments, cost was rarely mentioned explicitly. This contrasts sharply with sampling developments in the 1960s when the cost savings associated with conducting surveys by telephone led marketers and other non-governmental groups to move away from face-to-face household surveys. In this wave of change, government studies were not early adopters, at least partially because the sampling theory for efficient sampling of telephone numbers had not yet been developed. The U.S. government did not heavily use telephone sampling until the 1980s, well after statistical theory (Waksberg 1978) and standard practice for telephone surveys had been established in other sectors.

While telephone sampling was not a paradigm shift like the introduction of random sampling had been decades earlier, both developments provide some clues to the progression of ideas and practices. In both cases, statistical theory was not in the vanguard of the changes; rather, statistical theory was developed to support existing practices.

In today's environment, cost is probably the single most important factor driving the search for new methods of sampling. And the low cost of data collection on the Web is leading many, especially those in market research, toward this approach even without a widely accepted statistical theory to support it (Baker et al. 2010). This closely mirrors the adoption of telephone sampling in the 1960s. However, statistical theory and data collection practice eventually coalesced to provide a relatively solid foundation for telephone sampling. This combination is essential for a new methodology to be successful across a wide variety of applications. Marrying theory and practice remains a challenge today for new methods of sampling or surveying to take advantage of the low cost of the Web.

The role of new innovations in technology as a driver of survey methods is related to what Frankel and Frankel (1987) discussed regarding the influence of technology in the development of survey sampling. Bellhouse (2000) makes this interplay between survey sampling and technological innovation the focus of his history of sampling. There is little question that technology and survey sampling have been closely intertwined over the years, and each has at one time led to new developments in the other area. It is hard to imagine that the back-and-forth developments in sampling and technology will not continue in the future.

The explosive change in the way people communicate recently is a technological development that has given rise to a great deal of survey sampling research. One example is the transition to wireless or cell phones in the U.S. population. About 25 percent of adults lived in homes with no landline telephone in 2009 (Blumberg and Luke 2010). Accompanying this development is an impressive number of articles on cell phone sampling in the past five years (AAPOR 2010). The Web has been around longer, and there are even more articles and books on using the Web for surveying. Researchers have also begun to explore potential survey applications for texting, smart phones, and social networks. More developments in these areas are likely. Again, most of these technologies are being used at least partially due to their low cost, the main exception being cell phones, which are more expensive than landlines for surveying.

It is important to notice that these changes driven by technology are incremental and not paradigm shifts. It is fairly safe to suggest that the near-term future is likely to continue to see these types of developments, although major changes in the way we think about making inferences may not result from any of these developments.

Another low-cost alternative is the use of administrative records, although this approach is not new or driven by new technology. Despite their promise, the use of administrative records has not lived up to its potential, at least not in U.S. applications. Jabine and Scheuren (1985) describe six goals set by the U.S. government to improve the use of administrative records in survey sampling for a 10-year period. Roughly 25 years later, many of the goals outlined there have not been fully realized. Sampling from administrative records has not been hugely successful in areas other than government statistics either. For example, election surveys have examined the use of sampling from registration lists, but those efforts have encountered serious difficulties. Administrative records have always posed problems for survey applications due to incompleteness, lack of timeliness, poor data for locating and contacting the respondents, and incomplete maintenance of the records. These failings are so common because the purpose of the administrative records may not require the same level of quality as is needed for sampling purposes. Some of these problems are nicely captured in the National Research Council (2010) review of state voter registration databases. This review makes it clear why administrative records are not as useful as might be expected in one particular area. The future use of administrative records, while still promising, faces major challenges.

Much of the current research in sampling theory has been investigating estimation methods to account for nonresponse, coverage error, and other forms of missing data. This approach could also be considered a low-cost alternative, since it replaces expensive data collection with analytic methods or weighting adjustments. For example, adjustments are made to the base weights (the inverse of the probability of selection weights) to account for nonresponse, essentially transferring the weight from the nonrespondents to the respondents.

A seminal contribution in this area is Deville and Särndal (1992), who placed poststratification estimation methods into a class of so-called calibration estimators. Their article opened the floodgates to a new avenue of research, and calibration estimators were quickly generalized to deal with unit nonresponse and undercoverage—imperfections in the surveying process that are not directly handled by design-based theory. Rubin (1976) provided another boost to estimation research by suggesting multiple imputation as a method of variance estimation for imputed data. Imputation is a technique that is generally used to "fill in" or impute responses for items that are not completed by the respondent. Imputation for missing items has been around for decades, but multiple imputation spurred new research and attention. His work spawned thousands of articles in a variety of applications, including survey sampling.

The application of the idea of repairing shortcomings in data collection in probability samples using estimation techniques has also been explored for other sampling methods. Online surveys have used alternative estimation methods, especially propensity score weighting methods, to try to compensate for the non-random selection of participants (Lee and Valliant 2009; Schonlau et al. 2007). This is an area in which continued research is undoubtedly coming, and we discuss it in more detail below.

This short overview of the past and present suggests that cost is important, but was not the driver of innovations in sampling at the beginning of the era of probability sampling. Since then, cost has emerged as the primary agent for changes in sampling methods, even if these have largely been incremental changes. Statistical theory has seldom, if ever, been the leading agent of change. But statistical theory has been essential to supporting new developments. When statistical theory for a sampling method does not garner widespread acceptance for that methodology, then the sampling method is not likely to be accepted across disciplines and applications.

Before jumping ahead to the future, it is worthwhile to think more deeply about cost since it is so influential. In many ways, the increase in survey cost that has been experienced in recent decades is a symptom of the underlying problem of a decline in survey participation. Both telephone and face-to-face data collection would be considerably less expensive today than they were 20 or 30 years ago if sampled persons were as willing to participate as they used to be. As survey researchers, we often conclude that this lowered cooperation rate is solely a function of societal changes that we are powerless to influence. Although society has undoubtedly changed, survey researchers do have tools that might be effective in improving the quality of research and lowering cost. We believe that future research that focuses on the respondent rather than the analyst of the survey might help accomplish these twin objectives. A respondent-centric methodology was proposed by Dillman (1978). Respondent-centric methods attempt to increase cooperation to the survey request, thus reducing the need for costly data collection follow-ups and weight adjustments. The basic idea is that we should craft surveys so that the respondent perceives the benefits of participation as outweighing the

burdens of that participation. If this is possible, then both lower costs and higher quality can be achieved. More research in this area is needed.

Peering into the Mist

This perspective on the past and present implies directions that sampling might take in the future. In the near-term future, what Frankel and Frankel (1987) called innovations of the basic methods and extensions to accommodate new technologies is likely to continue at an accelerated rate. As an example, one area that is ripe for such developments is multiple frame sampling. In multiple frame sampling, the units are selected from two or more frames to increase coverage or efficiency. For example, in a survey of businesses, one frame might be an easily accessible list of businesses that is not complete, and the other frame might be a sample of geographic areas where businesses are listed and then sampled. Using the two frames provides more complete coverage while gaining efficiency by taking advantage of the available list.

Multiple frame sampling has even more promise because it offers a way of reducing costs by using administrative records. Enhancements in statistical theory as described by Lohr (2009) indicate that statisticians are interested in the topic. Statistical researchers are likely to continue to work on this problem because there are a number of unresolved technical issues, especially those associated with missing data due to the incompleteness of the records and measurement error.

In addition, multiple frame theory has recently been applied to cell phone surveys, and this research has revealed some basic deficiencies in sampling theory that might be addressed in the near-term future. For example, Brick (2010) shows that incorporating nonresponse at the design stage may result in both a different optimal allocation of the sample and a different weighting approach. Lohr (forthcoming) discusses how measurement error influences the choice of estimators in dual frame surveys.

The premise of sampling theory still assumes that coverage of the population in the frame is complete and that accurate responses will be obtained from all sampled units, despite the fact that these pristine conditions are virtually never encountered. Efforts to deal with these imperfections have almost totally relied on innovations in the estimation stage. An important new direction for the future of sampling is to consider the effects of (and attempt to ameliorate) nonsampling error in the sample design and data collection components rather than making estimation carry the burden alone. This expands upon the idea of responsive design (Groves and Heeringa 2006), which tries to reduce nonresponse bias in the data collection stage by targeting efforts to make the characteristics of the respondents more consistent with those of the sample. Similarly, the importance of strong auxiliary data has been identified as a key requirement for reducing nonresponse bias in recent years. Since sample design also benefits when such data are available, different methods for

incorporating these data at the sampling stage might be a new avenue of research. A change of this nature at the sample design stage would be a substantial shift that could happen in the near-term future.

Data collection cost is going to continue to force samplers to examine how they can take advantage of cheaper methods of data collection. A statistical theory that supports collecting observations from the Web from a nonprobability sample is an indispensible ingredient if we are to achieve this much-sought goal. An unresolved question is whether this goal can be accomplished within design-based probability sampling theory. If it is possible, then it is likely that sampling will be invigorated with many new applications and extensions. If not, two outcomes seem realistic: (1) a new paradigm could be introduced that accommodates Web surveys and this theory becomes generally accepted, replacing or supplementing design-based probability sampling; (2) collection of data from volunteers on the Web will be restricted to specific disciplines or applications because of the weak theoretical basis.

NONPROBABILITY SAMPLES

Whereas speculating about the nature of a new theory to replace probability sampling is too daunting a task, a critical review of approaches that are being investigated to make inferences from nonprobability or volunteer samples is feasible. The analysis of observational data is the statistical area that is most closely related to the analysis of volunteer Web samples. In fact, much of the research on Web samples has used techniques that were originally developed for observational studies, most notably propensity scoring. For this reason, we begin by discussing propensity scoring methods.

Propensity scoring is a technique proposed for analyzing observational data, or a set of data not selected based on randomized sampling (Rosenbaum and Rubin 1983). The basic idea is to compare outcomes from a "treated" group of individuals with those from an "untreated" group to assess the causal effect of the treatment when the units are not assigned to the treatments randomly. Propensity scores are the conditional probability that a unit is treated, given a set of auxiliary variables, under a model. If a researcher can balance the treated and untreated groups on the propensity scores and the model is correct, then the causal effect of the treatment can be estimated unbiasedly.

The appeal of this technique for volunteer samples is obvious; however, there are important differences between observational studies and surveys. One important difference is that most surveys are intended to address a host of diverse research objectives, even when one specific goal such as measuring the unemployment rate guides the design of the survey. Observational studies, on the other hand, are most often mounted to answer a small number of specific questions—with the primary goal of determining the effect of a treatment.

The researcher using propensity scores for an observational study aims to build a model that identifies pathways that lead to self-selection into the treatment (or to exclusion from the treatment). As daunting as this challenge is, the

observational data researcher needs to identify only variables for the model that are potentially related to the treatment effect. In a survey setting, the development of a propensity score model to account for self-selection is much more difficult because the scope is not a single treatment or outcome measure. Self-selection variables and variables related to multiple outcome variables must be included in the model appropriately to reduce bias.

Propensity scoring has been used in probability samples to compensate for nonresponse and undercoverage. When reasonable auxiliary variables are available in this type of application, propensity scoring (or another nonresponse adjustment method) has been shown to reduce very large biases substantially. These methods are less effective for smaller biases. Two reasons for the inability to consistently reduce bias due to nonresponse may be specification error (primarily the exclusion of important variables or failing to account for interactions between variables) and random measurement error. The specification problem is well known from regression analysis, where excluding predictors or failing to include interactions between variables has been shown to bias the estimates. Steiner, Cook, and Shadish (2011) discuss the bias that may arise due to random measurement error, and their findings are directly applicable to volunteer samples. They show that measurement error in the auxiliary variables has the potential to cause substantial bias in the estimates of the treatment effects even if the specification of the model is correct. Thus, propensity scoring and other estimation methods must not only correctly specify the models, but they must have auxiliaries with low measurement errors.

The adjustment techniques applied in probability samples also have a major advantage over the use of the same methods in nonprobability or volunteer samples. In sample surveys, all the sampled units are subject to a similar stimulus to encourage them to participate in the survey. Thus, the specification or modeling task for the propensity model should incorporate auxiliary variables related to the sampled unit's willingness to accept the offer to participate (since there are many outcome variables, only the most important of these may also be included). Even this task is complex. For example, we often see different relationships and variables being important depending on the reason for nonresponse. For example, noncontact and refusal cases may have different relationships with the auxiliaries because the stimulus may not be consistent. Lin and Schaeffer (1995) found that contact and refusal cases generated biases that were in the opposite direction, making the specification problem difficult.

Another advantage of the use of adjustment methods in probability samples is that nonresponse in surveys is a much studied phenomenon with a substantial body of knowledge. Groves and Couper (1998) summarize many of the known relationships between response and characteristics being estimated. This information can be very useful in building adjustment or propensity models.

With online volunteer samples, one of the complications is that all members of the target population do not have the same exposure to the invitation to the online survey. Some population members may not be accessible because they

are not online, while others may be inaccessible for other reasons (e.g., they never go to the Web pages that advertise the survey or they set technological barriers to avoid being exposed to these types of invitations). Others may be exposed but are not interested in participating for a variety of reasons (similar to the nonresponse problem of surveys). Much of the modeling of bias in Web surveys to date has focused on only one aspect of this underexposure, such as those who are not online. Trying to do a reasonable specification for a full set of pathways that might be important and correlated to a variety of outcome variables is exceedingly difficult.

It is interesting that quota sampling has not been mentioned thus far, even though quota sampling is the one method that resisted the probability sampling revolution in the first part of the twentieth century. Quota sampling was able to gain a foothold and retain this position because it often gave reasonable estimates and was less expensive than probability sampling. It is still popular in many countries. Stephan and McCarthy (1958) examined the empirical evidence and found that quota sampling was not consistently inferior to probability samples in terms of bias, although there were situations in which quotas increased the bias substantially. Today, quota sampling is still used in some circumstances in the United States, but many, including the federal government, do not accept quota samples when important actions may be taken as a result of the survey estimates. The loss of the popularity of quota sampling in the United States may be more a function of a few highly visible failures in election surveys where quotas were partially responsible for the biases in the predictions of winners and margins of victory. These highly publicized survey failures are more examples of where the methods of implementing surveys have caused public concern about their accuracy.

The major, but not the only, shortcoming of quota sampling is the bias resulting from the selection of respondents within the quota classes by interviewers. If we think about quota sampling in conjunction with propensity scoring, the analysis task would be to model the selection process controlled by individual interviewers. This would pose a formidable model building activity (one that quota samplers try to avoid by setting the quota groups or classes in the design stage). If the selection is removed from the interviewers, there is little cost savings or benefit relative to probability sampling.

Returning to online surveys, Rivers (2007) takes a different approach that has some similarities to quota sampling but is more analogous to case-control studies. Rivers suggests sample matching, where a "target" sample is selected from a list or frame, preferably using probability sampling. For example, a target sample might be randomly selected from a Census Bureau sample such as the American Community Survey or the Current Population Survey. The characteristics of each target sample unit are compared to those of members of Web panels, and the closest match in the pool of available respondents from the Web panel is interviewed. The Web respondent essentially is a substitute for the randomly sampled matched unit. Rivers requires a large set of relevant auxiliary

data for both the target and the Web panel to improve the quality (reduce the bias) of the interviewed sample.

This approach has some nice features and deserves further study. Like quota sampling, the probability of selecting a respondent cannot be computed. Unlike quota sampling, the criteria for choosing the matched unit are more detailed and are not subject to interviewer or other subjective methods that often increase biases. A weakness of the technique is that it is very dependent on the modeling assumptions linking the target sample to the Web panel members. This is also the primary weakness in case-control studies. If the outcome variables are homogeneous within the matching classes, then the inferences for those outcomes should be robust. Note that this is the same condition for quota samples to be unbiased, but in this case the auxiliaries may provide additional homogeneity as compared to the usual quota classes.

Another way of thinking about the matching approach of Rivers is that it is an efficient sampling technique. The matching is fine stratification and should lead to relatively precise estimates. However, the sample from the Web panel can be used to make unbiased estimates of the target population only if the outcome characteristics of the volunteers are equivalent to those of the target population conditional on the auxiliaries—a strong assumption.

The procedure depends on having powerful auxiliary variables to form matching classes and on a good understanding of the relationships of these auxiliaries with outcome variables of interest. These relationships are essential for the modeling activity. Based on the empirical findings from the 2010 elections where Rivers's approach was applied successfully, these conditions may exist for predicting elections. The biases for estimating other characteristics, especially those that have not been studied as closely as voting behaviors, are likely to be larger. The use of this approach for a wide variety of outcome variables is also very challenging.

IS IT STILL A PROBABILITY SAMPLE?

Some researchers have suggested that a probability sample with a low response rate is itself a volunteer sample, and therefore does not have any advantages over a nonprobability sample. Based on the arguments given above, we contend that a well-conducted probability sample with a low response rate is likely to be of higher quality (lower average bias for a broad range of estimates) than a sample of volunteers.

The proposition that response rates may be less problematic generally than the sampling approach is not an easily tested one. Empirical findings can only suggest whether this holds broadly. The basis for the proposition is that a probability sample has a much more uniform foundation to build upon since all the sampled units are actively recruited and encouraged to participate. Furthermore, large nonresponse bias results only under conditions that are less frequently

encountered in practice than generally recognized. For example, Groves, Presser, and Dipko (2004) conducted experiments intended to induce nonresponse bias so its properties could be studied. Despite their efforts to establish conditions that were likely to incur nonresponse bias, large nonresponse biases were the exception in their surveys. Large nonresponse biases do occur, but not as regularly as we might assume. Even when they occur, only a small subset of the estimates from the survey may have large biases.

It is not clear that the same benefits can be claimed for a probability sample drawn from a frame with serious coverage problems. Undercoverage is more insidious than nonresponse. The potential for bias with substantial undercoverage of the target population is in many ways more similar to the potential for bias in volunteer samples. The undercovered are not subject to the same stimulus to respond to the survey. Van de Kerckhove et al. (2009) reported that an RDD survey conducted in 2007 with response rates as low as about 40 percent had no evidence of large nonresponse bias for the set of estimates they evaluated. But they did find evidence of undercoverage bias due to the exclusion of households without landlines for the same survey. At that time, the coverage rate was above 80 percent. All missing data rates are not equivalent.

Concerns about biases in sample surveys with undercoverage may also have implications for the future use of administrative records for sampling. Administrative records are often incomplete and out of date, and typically the data are not missing at random. Sampling from incomplete administrative records may lead to undercoverage bias. Even when the administrative record database contains good auxiliary variables, the biases may persist for estimating a range of statistics. The reason is that the units in the target population that cannot be sampled because of coverage errors may not be easy to account for in the inference process because the mechanisms causing the missingness are often unknown and multifaceted. This is one of the reasons for beginning this section by suggesting that multiple frame surveys may be important in the future of sampling.

Random Samples of the Future

Reading the history of sampling reveals that the historical path is fragile, and that the course could have been very different if the conditions were slightly different. The more we look at the past and present events, the more timid we become in speculating about the course of the future. Baseball legend Yogi Berra is purported to have said, “The future ain’t what it used to be.” We couldn’t agree more.

One reason for the fragility of the past is the major role that individuals played in determining the path. Personalities were critical in deciding which ideas were adopted and which were not. As noted earlier, Kaier’s arguments before the ISI for a representative form of sampling were undertaken without

statistical theory and with limited empirical results. Yet he was able to convince an international audience of the importance of representative samples by repeated and persuasive presentations on the topic. Kruskal and Mosteller (1980) note that "Kaier's work in sampling was a one-man show." Individuals make a difference.

In the United States, the Census Bureau was an early adopter of probability sampling and Morris Hansen led many of the Bureau's efforts. Hansen greatly contributed to the theory and practice of sampling. He, like Kaier, was a proselytizer for the technique. He advocated for probability sampling in the United States and internationally. Everyone who had the opportunity to work with Morris would agree that he was an extremely talented leader, extraordinarily persuasive, and doggedly persistent. Without personalities like those of Kaier and Hansen, the history of sampling might have been very different. The future of sampling is also likely to be in the hands of personalities who have not yet been revealed.

Two other factors were important in the history of sampling and are likely to be relevant to its future. One is the wealth of scientific ideas circulating in the society, not necessarily coming from survey sampling or even statistics. The early twentieth century was a time of great development of scientific and statistical ideas. Neyman, whose 1934 article established a new paradigm for survey sampling, had only two contributions to survey sampling. He did, however, have an enormous influence on statistical theory in general with his many other contributions. In this type of rich environment, growth and change in methods is likely.

The second factor is the demand the society has for information. Survey sampling was able to flourish in part because twentieth-century society desired to know more about a wide range of topics and put the resources into these investigations. Obtaining observations from all the members of a population is too cumbersome and expensive to meet these demands. If the probability sampling paradigm had not emerged, we suspect that some other method or procedure for making inferences from a set of observations would have been developed to meet these needs.

One thing we are confident about is that the future of sampling will be dynamic. Our society has continually expanding needs, and the scientific progress in society is extraordinary. Survey sampling is surely going to undergo changes, and the changes might be larger in scope than anticipated here. While change is certain to occur, we are much less confident about the specific path the future will follow.

References

AAPOR Cell Phone Task Force Report. 2010. *New Considerations for Survey Researchers When Planning and Conducting RDD Telephone Surveys in the U.S. with Respondents Reached via Cell Phone Numbers*. http://www.aapor.org/AM/Template.cfm?Section=Cell_Phone_Task_Force&Template=/CM/ContentDisplay.cfm&ContentID=2818.

Baker, Reg, Stephen Blumberg, J. Michael Brick, Mick Couper, Melanie Courtright, J. Michael Dennis, Don Dillman, et al. 2010. "Research Synthesis: AAPOR Report on Online Panels." *Public Opinion Quarterly* 74:711–81.

Bellhouse, David. 2000. "Survey Sampling Theory over the Twentieth Century and Its Relation to Computing Technology." *Survey Methodology* 26:11–20.

Blumberg, Stephen, and Julian Luke. 2010. *Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, July–December 2009*. Hyattsville, MD: U.S. Centers for Disease Control and Prevention, National Center for Health Statistics. http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201005.pdf.

Bowley, Arthur. 1926. "Measurement of the Precision Attained in Sampling." *Bulletin of the International Statistical Institute 22:* Supplement to Liv. 1:1–62.

Brick, J. Michael. 2010. "Allocation in Dual Frame Telephone Surveys with Nonsampling Errors." Paper presented at the Statistical Society of Canada Annual Meeting, Quebec City, Canada.

Couper, Mick. 2000. "Web Surveys: A Review of Issues and Approaches." *Public Opinion Quarterly* 64:464–94.

Deville, Jean-Claude, and Carl-Erik Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87:376–82.

Dillman, Don. 1978. *Mail and Telephone Surveys: The Total Design Method*. New York: Wiley & Sons.

Duncan, Joseph, and William Shelton. 1978. *Revolution in United States Government Statistics, 1926–1976*. Washington, DC: U.S. Department of Commerce, Office of Federal Statistical Policy and Standards.

Frankel, Martin, and Lester Frankel. 1987. "Fifty Years of Survey Sampling in the United States." *Public Opinion Quarterly* 51(Part 2):S127–38.

Godambe, Vidyadhar. 1966. "A New Approach to Sampling from Finite Populations. II. Distribution-Free Sufficiency." *Journal of the Royal Statistical Society Series B (Methodological)* 28:320–28.

Green, Donald, and Alan Gerber. 2006. "Can Registration-Based Sampling Improve the Accuracy of Midterm Election Forecasts?" *Public Opinion Quarterly* 70:197–223.

Groves, Robert, and Mick Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: Wiley.

Groves, Robert, and Steven Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society Series A* 169:439–57.

Groves, Robert, Stanley Presser, and Sarah Dipko. 2004. "The Role of Topic Interest in Survey Participation Decisions." *Public Opinion Quarterly* 68:2–31.

Hansen, Morris, William Madow, and Benjamin Tepping. 1983. "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys." *Journal of the American Statistical Association* 78:776–93.

Heckathorn, Douglas. 1997. "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations." *Social Problems* 44:174–99.

Holt, D. Tim, and Smith. 1979. "Post-Stratification." *Journal of the Royal Statistical Society Series A* 142:33–46.

Jabine, Thomas, and Fritz Scheuren. 1985. "Goals for Statistical Uses of Administrative Records: The Next 10 Years." *Journal of Business & Economic Statistics* 3:380–91.

Kaier, Anders. 1895–96. "Observations et experiences concernant des dénombrements representatives." *Bulletin of the International Statistical Institute* 9:176–83.

Kalton, Graham, and J. Michael Brick. 1995. "Weighting Schemes for Household Surveys." *Survey Methodology* 21(1):33–44.

Keeter, Scott. 2006. "The Impact of Cell Phone Noncoverage Bias on Polling in the 2004 Presidential Election." *Public Opinion Quarterly* 70:88–98.

Kruskal, William, and Frederick Mosteller. 1980. "Representative Sampling IV: The History of the Concept in Statistics, 1895–1939." *International Statistical Review/Revue Internationale de Statistique* 48:169–95.

Lavalleće, Pierre. 2007. *Indirect Sampling*. New York: Springer.

Lee, Sunghee, and Richard Valliant. 2009. "Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment." *Sociological Methods & Research* 37:319–43.

Lin, I-Fen, and Nora Cate Schaeffer. 1995. "Using Survey Participants to Estimate the Impact of Nonparticipation." *Public Opinion Quarterly* 59:236–58.

Little, Roderick, and Donald Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: John Wiley & Sons.

Lohr, Sharon. 2009. "Multiple Frame Surveys." In *Handbook of Statistics, Vol. 29A, Sample Surveys: Design, Methods, and Applications*, edited by D. Pfefferman and C.R. Rao, 3–8. Amsterdam: Elsevier/North-Holland.

———. Forthcoming. "Alternative Survey Sample Designs: Sampling with Multiple Overlapping Frames." *Survey Methodology*.

National Research Council. 2010. *Improving State Voter Registration Databases: Final Report.* Washington, DC: National Academies Press.

Neyman, Jerzy. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97:558–625.

Rao, J. N. K. 2003. *Small Area Estimation.* New York: Wiley.

Rivers, Douglas. 2007. "Sample Matching for Web Surveys: Theory and Application." Paper presented at the 2007 Joint Statistical Meetings, Salt Lake City, UT.

Rosenbaum, Paul, and Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41–55.

Royall, Richard. 1970. "On Finite Population Sampling Theory under Certain Linear Regression Models." *Biometrika* 57:377–87.

Rubin, Donald. 1976. "Inference and Missing Data." *Biometrika* 63:581–92.

Schonlau, Matthias, Arthur van Soest, Arie Kapteyn, and Mick Couper. 2007. "Are 'Webographic' or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring?" *Survey Research Methods* 1:155–63.

Steiner, Peter, Thomas Cook, and William Shadish. 2011. "On the Importance of Reliable Covariate Measurement in Selection Bias Adjustments Using Propensity Scores." *Journal of Educational and Behavioral Statistics* 36:213–36.

Stephan, Frederick, and Philip McCarthy. 1958. *Sampling Opinions: An Analysis of Survey Procedure.* New York: Wiley.

Thompson, Steven, Fred Ramsey, and George Seber. 1992. "An Adaptive Procedure for Sampling Animal Populations." *Biometrics* 48:1195–99.

Thompson, Steven, and George Seber. 1996. *Adaptive Sampling*. New York: Wiley.

Van de Kerckhove, Wendy, Jill Montaquila, Priscilla Carver, and J. Michael Brick. 2009. *An Evaluation of Bias in the 2007 National Households Education Surveys Program: Results from a Special Data Collection Effort*. NCES 2009–029. Washington, DC: U.S. Department of Education, National Center for Education Statistics. http://www.nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009029.

Volz, Erik, and Douglas Heckathorn. 2008. "Probability-Based Estimation Theory for Respondent-Driven Sampling." *Journal of Official Statistics* 24:79–97.

Waksberg, Joseph. 1978. "Sampling Methods for Random Digit Dialing." *Journal of the American Statistical Association* 73:40–46.