REPORT OF THE AAPOR TASK FORCE ON NON-PROBABILITY SAMPLING

Reg Baker, Market Strategies International and Task Force Co-Chair
J. Michael Brick, Westat and Task Force Co-Chair
Nancy A. Bates, Bureau of the Census
Mike Battaglia, Battaglia Consulting Group, LLC.
Mick P. Couper, University of Michigan
Jill A. Dever, RTI International
Krista J. Gile, University of Massachusetts Amherst
Roger Tourangeau, Westat

June 2013

ACKNOWLEDGEMENTS

A number of individuals beyond the members of the Task Force made important contributions to this report by providing review and feedback throughout. They include:

Robert Boruch, The Wharton School of the University of Pennsylvania
Mario Callegaro, Google
Mitch Eggers, Global Market Insite
David P. Fan, University of Minnesota
Linda Piekarski, Survey Sampling International
George Terhanian, Toluna
Jan Werner, Jan Werner Data Processing
Clifford Young, IPSOS

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Survey researchers routinely conduct studies that use different methods of data collection and inference. Over about the last 60 years most have used a probability-sampling framework. More recently, concerns about coverage and nonresponse coupled with rising costs, have led some to wonder whether non-probability sampling methods might be an acceptable alternative, at least under some conditions.

There is a wide range of non-probability designs that include case-control studies, clinical trials, evaluation research designs, intercept surveys, and opt-in panels, to name a few. Generally speaking, these designs have not been explored in detail by survey researchers even though they are frequently used in other applied research fields.

In the fall of 2011 the AAPOR Executive Council appointed a task force "to examine the conditions under which various survey designs that do not use probability samples might still be useful for making inferences to a larger population." A key feature of statistical inference is that it requires some theoretical basis and explicit set of assumptions for making the estimates and for judging the accuracy of those estimates. We consider methods for collecting data and producing estimates without a theoretical basis as not being appropriate for making statistical inferences.

In this report, we have examined the strengths and weaknesses of various non-probability methods, considering the theoretical and, to some extent, empirical evidence. We do not claim to have produced an exhaustive study of all possible methods or fully examined all of the literature on any one of them. However, we believe that we have at least identified the most prominent methods, and examined them in a balanced and objective way.

Overview of Report

The report begins with a short introduction and background on the survey profession's use of probability and non-probability methods over the years. The goal of this review is to provide an idea of the evolution of the ideas that prompted AAPOR to convene this task force. In Section 3 we introduce some of the generic challenges of non-probability sampling, with a special focus on the difficulties of making inferences. We also describe some methods that we do not consider within the scope of the task force because there is no theoretical basis or no sample design component.

Sections 4, 5, and 6 describe in more detail the principal non-probability methods that survey researchers might consider. Each of these methods attacks the issues in somewhat different ways. One approach is sample matching, which has been used for observational studies for many years and has recently been advocated for use in surveys that use opt-in panels. A second approach is network sampling, including respondent driven sampling. RDS is increasingly used for sampling rare and hard to interview groups where probability sampling methods are often not feasible. The last of these three sections discusses a set of post hoc adjustments that have been suggested as ways to reduce the bias in estimates from non-probability samples; these adjustments use auxiliary data in an effort to deal with selection and other biases. Propensity score adjustment is probably the most well known of these techniques.

Sections 7 and 8 discuss methods for assessing the precision of estimates and the concept of fitness for use. Probability samples have a well-defined set of quality criteria that have been organized around the concept of Total Survey Error (TSE). Non-probability samples do not fit within this framework very well and some possible alternatives to TSE are explored. This probably is the greatest need if non-probability methods are to be used more broadly in survey research. The concept of fitness for use also is explored and seems to have great relevance for non-probability samples, as well as for probability samples. More development is needed in this area as well.

Conclusions and Recommendations

The final section presents the conclusions of the Task Force. Those conclusions are summarized below.

Unlike probability sampling, there is no single framework that adequately encompasses all of non-probability sampling. Non-probability sampling is a collection of methods and it is difficult if not impossible to ascribe properties that apply to all non-probability sampling methodologies.

Researchers and other data users may find it useful to think of the different non-probability sample approaches as falling on a continuum of expected accuracy of the estimates. Surveys at the lower and upper ends of the continuum are relatively easy to recognize by the effort associated with controlling the sample and post hoc adjustments. The difficulty arises in placing methods between these two extremes and assessing the risks associated with inferences from these surveys. The risk depends on substantive knowledge and technical features.

Transparency is essential. Whenever non-probability sampling methods are used, there is a higher burden than that carried by probability samples to describe the methods used to draw the sample, collect the data, and make inferences. Too many online surveys consistently fail to include information that is adequate to assess their methodology.

Making inferences for any probability or non-probability survey requires some reliance on modeling assumptions. Those assumptions should be made clear to the user and evidence of the effect that departures from those assumptions might have on the accuracy of the estimates should be identified to the extent possible.

The most promising non-probability methods for surveys are those that are based on models that attempt to deal with challenges to inference in both the sampling and estimation stages. Model-based approaches typically assume that responses are generated according to a

statistical model (e.g., the observations all have the same mean and variance). These models typically attempt to use important auxiliary variables to improve fit and usability. Once the model is formulated, standard statistical estimation procedures such as likelihood-based or Bayesian techniques are then used to make inferences about the parameters being estimated.

One of the reasons model-based methods are not used more frequently in surveys may be that developing the appropriate models and testing their assumptions is difficult and time-consuming, requiring significant statistical expertise. Assumptions should be evaluated for all the key estimates, and a model that works well for some estimates may not work well for others. Achieving the simplicity of probability sampling methods for producing multiple estimates is a hurdle for non-probability sampling methods to overcome.

Fit for purpose is an important concept for judging survey quality, but its application to survey design requires further elaboration. Organizations that conduct probability samples have attempted to balance quality characteristics including relevance, accuracy, timeliness, accessibility, interpretability, and consistency. A similar effort is needed for non-probability samples.

Sampling methods used with opt-in panels have evolved significantly over time and, as a result, research aimed at evaluating the validity of survey estimates from these sample sources should focus on sampling methods rather than the panels themselves. Users of opt-in panels may employ different sampling, data collection, and adjustment techniques. Research evaluations of older methods of non-probability sampling from panels may have little relevance to the current methods being used.

If non-probability samples are to gain wider acceptance among survey researchers there must be a more coherent framework and accompanying set of measures for evaluating their quality. One of the key advantages of probability sampling is the toolkit of measures and constructs (such as TSE) developed for it that provides ways of thinking about quality and error sources. Using that toolkit to evaluate non-probability samples is not especially helpful because the framework for

sampling is different. Arguably the most pressing need is for research aimed at developing better measures of the quality of non-probability sampling estimates that include bias and precision.

Although non-probability samples often have performed well in electoral polling, the evidence of their accuracy is less clear in other domains and in more complex surveys that measure many different phenomena. Surveys designed to yield only a handful of estimates on a related set of outcomes may require the control of only a small set of covariates. However, many surveys do not have these advantages. A survey often produces many estimates across a broad array of subject areas and domains, requiring a larger set of covariates.

Non-probability samples may be appropriate for making statistical inferences, but the validity of the inferences rests on the appropriateness of the assumptions underlying the model and how deviations from those assumptions affect the specific estimates. Throughout the report, we have emphasized the need for further development of a theoretical basis for any non-probability sampling method to be followed by empirical evaluation of that method. The evaluation should assess the appropriateness of the assumptions under various circumstances and for different estimates. Our review identified sample matching as one of method that already has a theoretical basis constructed for evaluation studies that could be modified and amplified for use with surveys. Several researchers have begun this effort already. The post-survey adjustment methods applied to non-probability sampling have largely mirrored efforts in probability samples. Although this may be appropriate and effective to some extent, further consideration of selection bias mechanisms may be needed. We believe an agenda for advancing a method must include these attributes.

1. INTRODUCTION

Survey researchers routinely conduct studies that use different methods of data collection and inference. Over about the last 60 years most have used a probability-sampling framework. More recently, concerns about coverage and nonresponse coupled with rising costs have led some to wonder whether non-probability sampling methods might be an acceptable alternative, at least under some conditions.

There is a wide range of non-probability designs that include case-control studies, clinical trials, evaluation research designs, intercept surveys, and opt-in panels, to name a few. Generally speaking, these designs have not been explored in detail by survey researchers even though they are frequently used in other applied research fields. Because of their limited use in surveys, the assumptions required to make valid inferences from non-probability samples are not well understood by survey researchers.

In the fall of 2011 the AAPOR Executive Council appointed a task force "to examine the conditions under which various survey designs that do not use probability samples might still be useful for making inferences to a larger population." We recognize that the term "statistical inference" has many definitions and meanings. In this report, we take it to mean a set of procedures that produces estimates about the characteristics of a target population and provides some measure of the reliability of those estimates. A key feature of statistical inference is that it requires some theoretical basis and explicit set of assumptions for making the estimates and for judging the accuracy of those estimates. We consider methods for collecting data and producing estimates without a theoretical basis as not being appropriate for making statistical inferences.

Some readers may expect this report to focus, at least partially, on comparisons between probability and non-probability methods, contrasting the strengths and weaknesses of each. Those readers likely will be disappointed. We have explicitly avoided exploring probability sampling methods under less than ideal conditions and comparing of estimates between probability and non-

probability samples. We realize there is considerable interest in whether a probability sample is still a probability sample when it has low coverage or high nonresponse, but the task force has not attempted to undertake this controversial and substantial task.

What we have done is examine the strengths and weaknesses of various non-probability methods, considering the theoretical and, to some extent, empirical evidence. We do not claim to have produced an exhaustive study of all possible methods or fully examined all of the literature on any one of them. However, we believe that we have at least identified the most prominent methods, and examined them in a balanced and objective way.

Non-probability sampling has become especially prevalent as more and more surveys have moved online. More often than not, the primary sample source for online research is a panel of individuals who have been recruited in advance and agreed to do surveys. In this report we use the term *opt-in panel* to mean an online panel not recruited via probability sampling. The sampling approaches used with these panels vary substantially. In recent years, researchers working with opt-in panels have begun to explore techniques that go beyond simple quota sampling. The key point is that opt-in panels are not based on a single sampling method, but rely on myriad and varied sampling methods. Evaluations of survey results that use them should focus more on the sampling method.

Finally, we recognize that for many in AAPOR the terms "scientific survey" and "probability sampling" are nearly synonymous. Although the same can be said for many members of this task force, we attempted to be open-minded and fair in our review. If you firmly hold that statistical inference is impossible without probability sampling, or if you firmly believe that the sampling method is irrelevant to inference, then this review is unlikely to have a great deal of value for you. In this regard readers might do well to recall these words from Kish (1965):

> Great advances of the most successful sciences - astronomy, physics, chemistry - were and are, achieved without probability sampling. Statistical inference in these researches is based on subjective judgment about the presence of adequate, automatic, and natural randomization in the population . . . No clear rule exists for deciding exactly when probability sampling is necessary, and what price should be

> paid for it . . .Probability sampling for randomization is not a dogma, but a strategy, especially for large numbers. (pp. 28-29)

The purpose of this report is not to specify a "clear rule." Rather, we mean it to be a first step that we hope stimulates a broader debate across the profession. We think such a debate is much needed, if not overdue, as we face the challenge of adapting our methods to an already much-changed and continually evolving world. It is not the first such challenge in the long history of survey research, nor will it be the last. And as always has been the case, the profession will be stronger by taking it on.

2. BACKGROUND

Probability sampling has been the dominant paradigm for surveys for many decades, but it has by no means been the only paradigm nor has it always been dominant. This is not the place for a full review of the history of survey sampling. Excellent overviews can be found elsewhere (e.g., Frankel and Frankel 1987; Brick 2011). In this section we review that history only briefly while noting some of the key elements in today's debate over the merits of probability-based versus non-probability-based designs.

The principal goal of survey sampling is to make reliable and accurate inferences to a broader population. This is often referred to as "representation" although, as Kish (1995) noted, the term lacks precise meaning. Neyman's (1934) paper generally is viewed as a turning point for survey sampling. Before that, two main approaches to conducting surveys were used (see, for example, Kiaer 1895-6 and Yates 1946). The first involved choosing a representative community (using judgment), and then conducting a census of that community. The second involved purposive selection, that is, choosing sampling areas and units based on certain criteria or controls, akin to that of quota sampling. Neyman's comments on a 1929 paper describing the use of purposive sampling in Italy (Gini and Galvani 1929) ring true in today's debates about non-probability sampling: "The comparison of the sample and the whole country showed, in fact, that though the average values of seven controls used are in satisfactory agreement, the agreement of average values of other characters, which were not used as controls, is often poor. The agreement of other statistics besides the means, such as frequency distributions, etc., is still worse" (Neyman 1934, p. 585). Neyman's paper not only cited the weaknesses in such contemporary practices as purposive sampling and full enumeration, it also laid out the ideas that form the basis for the theory and practice of scientific survey sampling. In Kish's words, Neyman established "the triumph of probability sampling of many small, unequal clusters, stratified for better representation" (Kish 1995, p. 9).

While this "triumph of probability sampling" was largely true of official statistics collected by national statistical agencies, the probability sampling paradigm did not fully enter the public polling realm until after the two spectacular public failures of non-probability based surveys in the elections of 1936 and 1948. In the case of the 1936 election, the outcome was incorrectly predicted based on the 2.3 million straw poll ballots returned out of 10 million distributed by the *Literary Digest* (see Bryson 1976; Squire 1988). A poll conducted by George Gallup using quota sampling correctly predicted the outcome. In 1948, the three leading pollsters at the time (Crossley, Gallup, and Roper) all used quota sampling methods, and all three incorrectly predicted the winner of that election.

An influential review by Mosteller and colleagues (1949) of the 1948 results identified many sources of error in the pre-election polls. One potential source was the use of quota sampling rather than probability sampling. They noted in their report that the use of quota sampling resulted in interviewers selecting somewhat more educated and well-off people, which biased the sample against Truman. Although the report does not blame quota sampling per se for the incorrect predictions, it does question its use primarily because it provided no measure for the reliability of the poll estimates. The consequences were evident as Gallup adopted probability-based methods soon thereafter, and other pollsters followed suit. The field in general came to accept the proposition that a small well-designed sample survey can yield more accurate results than a much larger, less controlled design.

Still, estimates from non-probability samples are sometimes accurate. It is often forgotten that the *Literary Digest* had conducted polls in every election from 1920 until 1936, and correctly predicted the winner in each of them. Yet one very public failure led, or at least contributed, to the demise of the magazine in 1938.

In their review of the history of survey sampling in the U.S., Frankel and Frankel (1987 p. S129) wrote, "After 1948, the debate between the advocates of quota sampling and probability sampling was over, and probability sampling became the method of choice in the United States." Nonetheless, quota sampling continued to be used alongside probability-based methods for several more decades,

especially in market research, where it continues to this day, and even in academia. For example, the U.S. General Social Survey (GSS) used quota sampling in the first few years before switching to random sampling in 1977. A combination of random sampling (for primary and secondary sampling units) and quota sampling (for households or persons within such units) was common in Europe for many decades, and is still employed in some countries (see, e.g. Vehovar 1999). Substitution for nonresponse also is still a relatively widespread practice outside North America (e.g., Vehovar 1995). Similarly, the notion of "representative sampling" as in choosing a community and studying it in depth with a view to making inference to the broader society continued to persist in Russia and other countries until fairly recently.

Historically, the main arguments advanced against probability-based samples have been those of cost and time efficiency. This was an easy argument to make when the most common survey method was face-to-face. The introduction of random digit dial (RDD) telephone surveys in the 1970s (see Glasser and Metzger 1972) changed that equation. Extending the probability-based method to telephone surveys helped overcome the coverage problem of directory samples and the substantial effort required to draw and interview such samples. With RDD, surveys could be done relatively inexpensively and quickly with reasonable coverage of the full population and, in the early days at least, relatively low nonresponse.

The emergence of RDD heralded a broad expansion of survey methods in political polling, market research, and academia. This boom continued until the rapid rise in cell phone only households (Lavrakas et al. 2007) raised concerns about coverage bias. At the same time, the long-term decline in response rates, brought on in part by broad use of the telephone both for survey research and especially for telemarketing, raised questions about nonresponse bias (see, e.g., Brick and Williams 2013; Curtin, Presser, and Singer 2005).

The combination of rapidly increasing costs associated with the traditional probability-based methods (face-to-face and telephone), declining response rates, and rising concerns about telephone

survey coverage raised expectations about the potential benefits of online surveys, especially as Internet penetration increased (Couper 2000). However, the inability to develop RDD-like methods to sample and recruit respondents to web surveys led to the development of alternative approaches relying on non-probability methods, most notably opt-in panels comprised of volunteers. These panels offered the promise of yielding survey data from large numbers of respondents in very little time and at a relatively low cost for data collection. With access to millions of potential respondents, subgroups or those with special characteristics identified in profile surveys could be selected for specialized surveys. The early arguments in favor of using these panels for inference were based on their size (reminiscent of the *Literacy Digest* arguments), higher response rates (at least in the early days) than those being achieved in many telephone surveys, and on the ability to collect auxiliary data for adjustment. The popularity of these panels, not only for market research but also for political polling and even some academic research, caused the industry to take a close look at the reliability of results from such panels (these issues are thoroughly reviewed in Baker et al. 2010). However, as the AAPOR Task Force on Online Panels noted, such surveys certainly have value for some types of research, but researchers "should avoid nonprobability opt-in panels when a key research objective is to accurately estimate population values …claims of 'representativeness' should be avoided when using these sample sources."

And so the survey profession once again faces a significant challenge. As Frankel and Frankel (1987, p. S133) noted: "Prior to 1960 refusals were not regarded as posing a threat to inferences from sample surveys. The not-at-home respondent was the main problem, but this could be resolved by making repeated callbacks." In the last ten years or so, increasing effort (i.e., more callbacks) has not been sufficient to stem the tide of nonresponse, nor is it financially sustainable. The issue of low response rates to probability-based samples and related concerns about nonresponse bias are at the heart of the arguments in favor of alternative approaches.

This is not to say that alternatives such as automated telephone surveys (robo-calls) or opt-in web surveys do not suffer from nonresponse or coverage problems of their own. The proponents of these approaches are instead arguing that the traditional practice of probability sampling, in which good coverage of the general population and high response rates are seen as essential to minimizing inferential risk, is now so difficult and costly to achieve that the method may only be feasible for very well-funded and socially important surveys such as those done by national statistical agencies. They further argue that when the proper methods and variables are used to adjust the set of respondents (however obtained) to match the population of interest, valid inference is possible. This debate is often characterized as the difference between a design-based approach and a model-based approach.

Groves (2006) also addressed the nonresponse issue in a provocatively-titled section of his paper, "With high nonresponse rates, why use probability sampling?" He noted that non-probability designs burden the analyst with adjusting the respondent estimate both for the nonrandom selection and for nonresponse. For designs – such as access panels – that restrict those who may volunteer to a subset of the population, concerns about coverage are also salient. As Brick (2011) recently noted, "Nonresponse, incomplete coverage of the population, and measurement errors are examples of practical issues that violate the pure assumptions of probability sampling."

These papers make the point that just because a survey is based on a probability sample, does not mean it is a valid and reliable reflection of the population it purports to measure. In the same way, just because a survey is based on self-selected methods does not automatically disqualify it from attention or invalidate its findings.

There have been a number of reported instances where non-probability samples have yielded results that are as good as or even better than probability-based surveys when measured against an external criterion. Most notably, these have been in the area of pre-election polls (Abate 1998; Snell et al. 1998; Taylor et al. 2001; Harris Interactive 2004, 2008; Twyman 2008; Vavreck and Rivers 2008; Silver 2012). Similarly, there are claims that alternative methods such as sample matching can be as

accurate as probability samples when sample matching is used (e.g., Rivers 2007), when the appropriate variables are used in propensity score adjustment (e.g., Terhanian and Bremer 2012) or when the assumptions of respondent driven sampling are met (e.g. Heckathorn 1997). In theory, if the assumptions are fully met – as with probability-based methods – the resulting estimates are expected to be unbiased.

In summary, the debate about the value of nonprobability surveys for broad population inference, and the tradeoff of cost and timeliness versus quality, is not a new one. Technological advances (particularly the Internet) are leading to the development of new methods (as foreseen by Frankel and Frankel 1987), and the volume of surveys has increased, fueling the debate. However, the issue itself has been with us in various forms since the early days of survey sampling.

Essentially, the argument is one about risk and the confidence that can be placed on the findings of a survey. Design-based approaches, while using models to adjust for undercoverage and nonresponse, provide some protection against the risk of sampling bias. Non-probability approaches rely more heavily on the appropriateness of the models and, in most cases, on the selection, availability and quality of the variables used for respondent selection and post hoc adjustment.

Of course, surveys are not the only form of inference and, unlike probability sampling, there is no single framework that encompasses all forms of non-probability sampling. In the sections that follow, we review the use of different non-probability methods in a variety of settings. But it is the sample survey (or its alternatives) that explicitly claims to make broad descriptive and analytic inferences to the larger population, and hence the debate about inference and probability-based versus non-probability sampling methods is felt most keenly in this area.

3. INTRODUCTION TO NON-PROBABILITY SAMPLING

In the previous section we noted that unlike probability sampling, which has a unified framework for sampling and making inferences to a population, no single framework encompasses all forms of non-probability sampling. Rather, there is a broad range of sampling methods, some of which are complex and sophisticated while others are simple and straightforward.

We take the view in this report that for a sampling method to be of value it must be capable of making a statistical inference from responses from a sample to a larger target population. It must have a set of procedures to produce estimates about the characteristics of the target population and provide some measure of the reliability of those estimates. (For example, in probability sampling we can estimate the mean or total of the target population and produce confidence intervals that provide information on the reliability of the estimates.) It therefore is essential that a method have some theoretical basis or explicit set of assumptions for making the estimates and for judging the accuracy of those estimates. Of course, every framework for making inferences, including probability sampling, makes assumptions that are not fully satisfied in practice; the extent of the deviations from the assumptions are critical to whether the statistical inference is useful. We consider methods for collecting data and producing estimates without a theoretical basis as *not* being appropriate for making statistical inferences. Convenience sampling is one such method. We describe convenience sampling here briefly for completeness, but due to the lack of theory we do not pursue it further in this report.

3.1 Convenience Sampling

Within many social science disciplines, convenience sampling is the predominant method of sample selection. For example, although psychologists sometimes use data from nationally representative probability samples, it is far more common for their studies to be based on convenience samples of college students. In the mid-1980s, David Sears (1986) raised concerns about this. He examined the papers published in the three leading social psychology journals of the time and found that "in 1980,

75% of the articles in these journals relied solely on undergraduate subjects… most (53%) stated that they used students recruited directly from undergraduate psychology classes". The picture had not changed when he examined papers in the same journals published five years later. Although this reliance on unrepresentative samples may have its weaknesses, as Sears argued, psychologists have continued to use samples of convenience for most of their research. Sears was concerned more about the population from which the subjects in psychology experiments were drawn than about the method for selecting them, but it is clear that even the population of undergraduates is likely not represented well in psychology experiments. The participants in psychology experiments are self-selected in various ways, ranging from their decision to go to particular colleges, to enroll in specific classes, and to volunteer and show up for a given experiment.

The use of convenience samples is hardly restricted to psychology (e.g., see Presser, 1984). Some forms of litigation research rely heavily on mall-intercept samples (Diamond, 2000); these are popular in trademark infringement cases, where the population of interest consists of potential buyers of a product. Malls are a convenient place to find members of the relevant populations for many such cases. Randomized trials in economics and education also make broad use of non-probability samples.

A good deal of medical research is also based on non-probability samples of convenience -- often patients to which the investigators happen to have ready access. Couper (2007) cited a number of health studies that used web samples to study conditions ranging from social anxiety disorder to ulcerative colitis; almost all of these were samples of volunteers. And, finally, many survey researchers who also use probability samples nonetheless rely on convenience samples for such purposes as focus groups or cognitive testing of questionnaires. So, the use of convenience samples is widespread, even among researchers who acknowledge the superiority of probability sampling in other contexts.

Definition of Convenience Sampling. Most sampling textbooks do not offer a formal

definition of convenience sampling, but instead lump this method of sampling together with other non-probability methods. So let us begin by offering this definition: *Convenience sampling* is a form of non-probability sampling in which the ease with which potential participants can be located or recruited is the primary consideration. As the name implies, the participants are selected based on their convenience (for the researchers) rather than on any formal sample design. Some common forms of convenience sampling are mall-intercept samples, volunteer samples, river samples, samples for some observational studies, and some snowball samples. We briefly discuss each of these types of convenience sample in turn, although it is worth noting that in a few applications these methods may result in samples that are not convenience samples as defined.

Mall Intercepts. In a mall-intercept survey, interviewers attempt to recruit shoppers or other passersby at one or more shopping malls to take part in a study. Generally, neither the malls nor the individual participants are selected via probability sampling, although some systematic method may be used to determine who gets approached (or intercepted) within a given mall. For example, interviewers might approach every n^{th} person passing a specific location in the mall. However, most mall-intercept surveys emphasize getting respondents quickly and cheaply (and yet with some appearance of objectivity). As a result, the selection of malls and the selection of individuals are often done haphazardly with little regard for the assumptions necessary to justify inference to a larger population. In litigation studies, it is generally essential that the respondents who are recruited for the study actually belong to the target population of interest so there may be detailed screening questions to establish their eligibility (to identify, for instance, potential buyers of some specific type of product). Although mall-intercept samples may not allow the researchers to make quantitative inferences to the larger population that the sample is meant to be represent, the courts generally see them as having greater value than the alternative — a few handpicked witnesses who testify about their reactions to the stimuli of interest.

Panels of Volunteers. As we have already noted, volunteer samples are ubiquitous in social

science, medical, and market research settings. Generally, the volunteers sign up for a single study, but in some cases they join a panel for some period and are then asked to take part in multiple studies. Consumer panels for market research have been around for at least 50 years, originally as mail panels (Sudman and Wansink 2002). More recently, numerous opt-in web panels have been recruited to complete online questionnaires with panel members receiving invitations to complete large numbers of surveys each month (Couper and Bosnjak 2010). These opt-in panels were the subject of an earlier AAPOR task force report (Baker et al. 2010) that generally found inferences from these sample sources to be less accurate than those from probability samples.

With the passage of time researchers who rely on opt-in panels have come to recognize the shortcomings in these sample sources. A growing body of work is now focused on methods for correcting the potential biases in these panels as a way to improve their accuracy and usefulness. We discuss these methods in some detail in Sections 4 and 6 of this report.

River Samples. River sampling is a web-based opt-in sampling approach to creating a sample of respondents for a single survey or a panel for repeated surveys over time (GfK Knowledge Networks 2008, Olivier 2011). River sampling most often recruits potential respondents from individuals visiting one of many websites where survey invitations have been placed. Attention catching techniques such as pop-up boxes, hyperlinks and banners are used to attract individuals to encourage these visitors to complete a survey or even join an opt-in panel. There are two selection aspects of river sampling.

First, one must decide which websites are appropriate to serve as clusters. These clusters are different from the use of clusters in a probability sample design where units in the population are uniquely associated with a single cluster. The website is a cluster from which visitors are recruited. A river sample based on a single website likely will yield a sample of individuals who are similar (i.e., homogeneous) on various demographic, attitudinal, and other factors. This should be avoided if broader representation is desired and, in practice, river sampling typically uses many websites. For a

survey of the entire adult population, a reasonably large number of websites might be used and this might alleviate the need to take account of the nature of the visits to each site and still obtain a heterogeneous sample. For a survey intended to cover a specific population, external information can be helpful in deciding which websites to use. In theory, one might draw a stratified sample of websites to improve coverage of the target population. In practice, sites generally are selected to optimize the tradeoff between cost and the expected yield of potential respondents, although demographic targeting is sometimes used as well.

Second, individuals willing to participate may need to be screened to see if they qualify for the survey. A wide variety of screening characteristics may be used when the objective is to have the respondent complete a specific survey. When the objective is to enroll the individual in a panel, he or she may be asked to complete an even larger profile survey.

Thus, river sampling is an opt-in web-based sampling technique that historically has been a form of convenience sampling. More recently, there has been increased emphasis on the use of more formal designs in sample selection with the goal of improving representativeness.

Observational Studies. Observational studies are studies that attempt to test hypotheses (often causal hypotheses) about medical or social phenomena without carrying out controlled randomized experiments. (We discuss observational studies in more detail in the next section.) Many observational studies also use volunteer samples although some are based on probability samples.

Consider the famous Framingham Heart Study, a landmark attempt to study cardiovascular disease (CVD) in a "representative sample" and to investigate the role of cholesterol, smoking, and exercise in the development of CVD. The town of Framingham, Massachusetts was selected largely for reasons of convenience (the town was approximately the right size, it had just one hospital, and it maintained a directory of the residents). The residents who joined the panel included a mix of volunteers and persons selected using systematic sampling from the town's census. Thus, the sample design was something of a hybrid, with the site selected based on convenience and some of the

individual participants selected through probability methods and the rest consisting of self-selected volunteers added to boost the overall sample size (see Dawber, Meadors, and Moore 1951, for an early description of the study, then still in the planning stages). Many clinical trials use a similar blend of random and non-random selection, with hospitals or practices selected non-randomly and individual patients selected through some combination of probability and non-probability methods.

Snowball Sampling. A final type of convenience sample is worth distinguishing — snowball samples. Snowball sampling began as a method for sampling networks and was not originally a form of convenience sampling. Coleman (1958-1959) was a pioneering user of the technique as a method for studying a person's social environment. For example, a person might be asked to name his or her best friend, who would then be interviewed and asked about his or her best friend. Goodman (1961) showed that a rigorous version of this method using a probability sample, which he called "snowball sampling," had desirable statistical properties. Later extensions of this method, however, typically used non-probability sampling variants of snowball sampling to find members of hard-to-reach or hidden populations. In many of these later applications, the snowball sample started with a sample of convenience (the *seeds*) from the population of interest, not with a random sample (as Goodman had intended). Still later, Heckathorn (1997) introduced a specialized variant of snowball sampling called respondent-driven sampling. This approach allows for specialized sampling and substantial assumptions to allow for estimates that are approximately unbiased. We discuss respondent-driven sampling (RDS) in more detail in Section 5 of this report, along with other forms of snowball or network sampling.

3.2 Obstacles to Inference

As we noted at the outset convenience samples are but one of many forms of non-probability sampling and are characterized by the ease with which a sample can be drawn. Like all non-probability samples, convenience samples are subject to several potential sources of bias, but unlike other non-probability methods their practitioners generally avoid serious attempts to correct that

bias. Still, they are illustrative of the obstacles to inference in non-probability sampling in general.

Consider the overall population of interest for the survey, or its intended target population. In the Framingham Heart Study, for example, the researchers intended to characterize all adults aged 30 to 50 as of January 1, 1950 or at least all American adults in that age range; that is, the target population was presumably not restricted to Framingham residents. However, there is a large disparity between the target population, on the one hand, and the population actually exposed to recruitment, on the other. This is a general problem with most non-probability samples; the population subject to recruitment is likely to be a small and unrepresentative portion of the target population of interest. (Probability samples can have an analogous problem in that the sampling frame may not offer complete coverage of the target population — that is, there is often some risk of coverage error even in a high quality probability sample — but the proportion of the target population that is omitted is likely to be much smaller than with a convenience sample.) With non-probability samples, it may be better to call this problem "exclusion bias" rather than "coverage bias," since the vast majority of the target population is likely to have no chance of inclusion in the sample.

A second issue is that non-probability samples often consist of volunteers and these volunteers may not be very representative even of the population that was exposed to recruitment, let alone of the larger target population. With a probability sample, the selection probabilities are determined by the researchers and can be incorporated into the estimation process (via weighting). With volunteer samples, however, the participation probabilities are determined by the volunteers and are effectively unknown. The likelihood of selection bias — reflecting systematic differences between the volunteers and non-volunteers on the variables of interest — may be substantial (Bethlehem, 2010).

The final inferential obstacle with non-probability samples is that participation rates (conditional on being recruited for the study) are often quite low. Studies based on opt-in panels cannot report proper response rates; instead, they should report *participation rates*, defined as the proportion of those panelists who were invited to take part in the study who ultimately complete the survey. Web

surveys based on opt-in panels often have participation rates in the single digits.

Although all three problems may not apply to all non-probability samples there are many instances in which they do. Consider a sample of panel members drawn from an opt-in panel and asked to complete a specific web questionnaire. The members of the panel may have been recruited from a relatively small number of web sites, effectively excluding the majority of whatever target population is of interest to the researchers and particularly those without Internet access. Only a small fraction of those who received an invitation to join the panel may decide to opt in, and only a small fraction of the panelists who are invited to complete the specific survey may take part. Thus, the final set of responses may be subject to large exclusion, selection, and non-participation biases. And, as we noted earlier, with volunteer samples, the completion probabilities needed to make adjustments are difficult to estimate and therefore difficult to incorporate in a weighting scheme. The post hoc adjustments that are made are based on a comparison of the achieved characteristics and the expected ones, not on completion probabilities.

3.3 Estimation

Non-probability samples are sometimes used to make population estimates, but they are often used for other purposes as well. For example, the vast majority of psychological experiments are not used to estimate a mean or proportion for some particular finite population — which is the usual situation with surveys based on probability samples — but instead are used to determine whether the differences across two (or more) experimental groups are reliably different from zero. In still other cases, no quantitative conclusions are drawn (e.g., when a sample is recruited to take part in a focus group).

Regardless of the intended use of the data, some assumptions are required in order to make statistical estimates and to assess the variability of those estimates. Too often researchers who use non-probability samples draw quantitative conclusions that treat the data as though they came from a simple random sample. This treatment greatly simplifies the analysis of the data, the computation of

standard errors, and the conduct of significance tests. It assumes that the method of selecting respondents is ignorable. The validity of the assumption that the method of sampling is not important for the analysis stage is a topic of considerable debate.

Correction Procedures. Rather than ignore the sampling and response mechanisms, some researchers attempt to compensate for the potential exclusion, selection, and non-participation biases by using any of several weighting procedures. For example, weights may be applied to bring the data from the opt-in panel into line with known population totals, thereby at least attempting to account for the underrepresentation of some population groups and the overrepresentation of others. A popular method for accomplishing this is *post-stratification* (Kalton and Flores-Cervantes 2003), discussed in more detail in Section 6. Post-stratification adjusts the sample weights so that the sample totals match the corresponding population totals in each cell formed by cross-classifying two or more categorical auxiliary variables. For example, the sample weights may be brought into line with population figures for each combination of sex, region, and age category. With probability samples, the case's initial weight typically is the inverse of its selection probability, which then is multiplied by an adjustment factor. Separate adjustments are done for each cell (although cells may be combined to avoid extreme adjustments). With opt-in panels, the initial weights are sometimes just set to 1. The population total is often only an estimate based on a large survey, such as the American Community Survey (ACS) or the Current Population Survey (CPS). Post-stratification will eliminate the bias due to selection or coverage problems if, within each adjustment cell, the probability that each case completes the survey is unrelated to the case's value on the survey variable of interest. Another way of stating this assumption is that the participants and non-participants in a given cell have the same distribution on the survey variable. This condition is sometimes referred to as the *missing at random* assumption (Little and Rubin, 2002).

A method called sample matching attempts to select a web sample that matches a set of target population characteristics from the outset (Rivers and Bailey 2009) rather than making an adjustment

to bring the two into alignment after the fact. For example, when a subsample of panel members is selected to receive an invitation to complete a particular survey, the subsample is chosen to match the composition of the population of interest on some set of auxiliary variables. Those auxiliary variables may be standard demographics (say, region by sex by age category) but also include attitudinal measures such as political ideology or openness to innovation. Differences between the sample make-up and the make-up of the population that remain could potentially be corrected for via propensity weighting (discussed in Section 6), at least to some extent. In theory, sample matching would have an impact on bias similar to the impact of post-stratification, since a matched sample starts by agreeing with a set of population figures and a post-stratified sample ends up by agreeing with them. Still, non-response can knock the responding sample out of alignment from the population figures and a matched sample may still require further post-survey adjustments (as Rivers and Bailey discuss). Sample matching is covered in greater detail in Section 4.

Effectiveness of Adjustment Procedures. Although in principle these methods of adjustments could work, removing some or all of the biases from the estimates derived from a non-probability sample, the key issue is how effective they are in practice. At least eight studies have examined this issue (see Tourangeau, Conrad, and Couper, 2013, for a review).

The studies evaluating the effectiveness of the adjustment procedures all used similar methods. The researchers started with the results from a web survey or they simulated the results from a web survey by examining the subset of respondents to a survey done by telephone or face-to-face who have Internet access. Then they compared the estimates from the real or simulated web survey to some set of benchmarks. The benchmarks may have come from a calibration study (a parallel study done by telephone or face-to-face with a probability sample); they may have come from the full sample, if the web survey results were actually based on the subset of respondents with Internet access; or the benchmarks may have come from outside survey estimates (such as the CPS or some other trusted source; see Yeager et al. 2011).

Despite the differences across studies in how they obtained their benchmarks and in which specific adjustment strategies they examined, the studies generally reported the adjustments seem to be useful but offered only a partial remedy for these problems. It is not clear whether weighting adjustments or related procedures can allow researchers to make accurate population estimates based on non-probability samples, such as opt-in web panels. Nonetheless, research continues with a particular focus on methods to identify a broader set of adjustment variables than those used in the reviewed studies. And, as we discuss in Section 7, there may be instances in which researchers can live with biases, provided they aren't too large for the purpose at hand.

Estimation based on data from experiments. As we noted, many experiments done by psychologists are carried out on convenience samples and this is true of many methodological experiments as well. In experiments, the key issue is whether two or more groups differ from each other on one or more outcome variables. That issue is typically examined statistically by testing whether some model parameter (such as the difference between the group means) is significantly different from zero. The statistical rationale for the significant test rests on the assumption that the participants have been randomly assigned to one of the experimental groups. As we noted earlier, researchers may then be tempted to infer that any statistically significant differences apply not only to the subjects in the experiment but also to a broader population. However, the experimental data do not provide support for this inference unless a random sample of the broader groups is selected prior to the random assignment. This is rarely done. As a result, researchers tend to be more concerned with the comparability of the experimental groups that they can control than how well they represent a broader group. In the classic terminology introduced by Campbell and Stanley (1963), researchers conducting experiments generally give more weight to the internal validity of the experiment than to its external validity or generalizability.

Is this disregard for sampling considerations reasonable? In other words, is it safe to assume that the effects of a treatment on the participants will be the same in the target population? It depends.

The results from an experiment are likely to have external validity when either of two conditions is met. First, it could be that the biases due to the use of a convenience sample are small. Psychologists sometimes argue that with very basic psychological processes, such as those involving memory or perception, the estimates from experiments are unlikely be affected much by selection bias. People are people and, with respect to some processes, their similarities are greater than their differences. However, recent studies suggest that this argument has been over-used and there is much evidence of variation in what are presumed to be universal attributes (Henrich, Heine, and Norenzayan 2010). Second, it could be that the biases are large, but that they more or less cancel out so that the difference in the biases is close to zero. Again, the rationale for this assumption is often unclear.

Still, a common concern is that the difference in bias across experimental groups is large and that it produces an overestimate of the experimental effect in the population of interest. For example, an experiment involving, say, alternative grid designs in a web survey may produce a significant effect in an opt-in web panel, but that effect might be considerably smaller among the general population, whose members have less experience with web surveys than most web panelists and who are therefore less sensitive to subtleties of web survey design. Even worse, it could be that the difference in biases is both larger than and opposite in direction from the differences in the population means; this would mean the conclusion drawn from the experiment was not just an over- or underestimate, but wrong — within the target population of interest, the mean for the treatment group is not larger (or smaller) than the mean for the control group but smaller (or larger). It is not clear how often this situation arises in practice.

3.5 Inference without Sampling

The last several years have witnessed the emergence of a number of interesting and innovative concepts that make use of naturally occurring and readily available data to measure population characteristics or predict future behavior. In general, they differ from the other methods described

in this report in two important ways. First, they often don't involve sampling at all but rely instead on amassing large amounts of data with the implicit assumption that large numbers reduce any potential bias. Second, they tend to avoid surveys or direct questioning of respondents about their attitudes and behaviors to avoid the data collection costs, instead trying to infer these attributes in other ways.

The techniques can be grouped into three broad categories: social media research, wisdom of crowds and big data. We discuss each briefly below although without serious evaluation since in our judgment they fall outside of any defined sampling framework.

Social media research. This family of methods uses a technique sometimes referred to as *web scraping* (Poynter, 2010) to harvest user generated content across the Internet from social networking sites, blogs, microblogs—any site where people express their opinions or document their behavior (See Schillewaert, De Ruyck and Verhaeghe, 2009). The unit of analysis is a verbatim comment rather than a sampled individual and datasets consist of large amounts of textual data. These datasets are processed using natural language processing software capable of categorizing text and classifying it according to sentiment—positive or negative and sometimes intensity. Unlike traditional survey research, generally no connection is made by the analyst between a bit of text and the characteristics of the individual posting it.

These techniques are increasingly used by companies to monitor the public perception of their products and services as well as to engage directly with their customers by responding to complaints on these same platforms. There also are attempts to use these data to predict behavior. For example, researchers at HP Labs have developed a technique that uses Twitter to predict movie box office receipts (Asur and Huberman, 2010). In another Twitter-based research study, O'Connor and his colleagues (2010) showed how closely tweets correlate with electoral and other public opinion polls.

Although not social media data per se, Google has demonstrated how aggregations of search

data can be used track trends that may reflect public opinion and behavior. In 2008 the *New York Times* (2008) reported, "There is a new common symptom of the flu, in addition to the usual aches, coughs, fevers and sore throats. Turns out a lot of ailing Americans enter phrases like 'flu symptoms' into Google and other search engines before they call their doctors." The reference is to a project called Google Flu Trends (www.google.org/flutrends) which attempts to show that the volume of searches on flu symptoms is highly correlated with reports of flu collected by the Centers for Disease Control (CDC). Because the Google data are real-time and the CDC data have a reporting lag the search data may be useful in predicting flu outcomes across geographic areas.

Wisdom of Crowds. The reference here is to a 2004 book by James Surowiecki, *Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. The central thesis is that a diverse group of deciding individuals is better at solving problems and predicting outcomes than are experts. Each individual's estimate is considered to have two components: information and error. The error term across the entire group is assumed to sum to zero. Using the average of the estimates is thought to give a usable estimate of some future outcome, such as the result of an upcoming election (see, for example, *The Telegraph*, 2012). Surowiecki postulated that the technique works best if the crowd is diverse and opinions are gathered independently of one another, but the book is short on specifics and theory as to why the method should produce accurate estimates.

The basic thrust of the wisdom of crowds approach has long been practiced in prediction markets such as those designed to predict election outcomes. One example is the Iowa Electronic Markets run by the Tippie College of Business at the University of Iowa (Intrade, 2012). Prediction markets typically allow members to buy and sell contracts based on externalities such as economic indicators or election results. The basic idea is to predict an outcome. So rather than using a member's response on for whom she plans to vote, prediction markets use the member's preditcions of who will win the election. Essentially, the members are placing a bet on an outcome, for example,

whether President Obama will win or lose the 2012 election. Rothschild (2009) and Erickson and Wliezen (2008) have examined the accuracy of prediction markets, finding mixed results.

Big Data. The term *big data* is increasingly used to describe a dramatic expansion of what many of us are used to calling administrative data. Big data refers to the vast amount of organized external data routinely generated by industry, business, and government. In the past these data often have been used in probability-based surveys. For example, customer lists typically are used as sampling frames for customer satisfaction studies. These same data also can be used to provide covariates for unit nonresponse adjustments and propensity score models.

More recently, some also have argued that big data can be used as an alternative to a probability sample. One prime example is the Longitudinal Employer Household Dynamics (LEHD) program at the U.S. Census Bureau. This is a voluntary partnership between state labor market information agencies and the U.S. Census Bureau to develop new information about local labor market conditions at low cost, with no added respondent burden. The program provides state and local area statistics on employment, job creation, turnover, and earnings by industry, age and sex of worker. However, because the raw data are based on Unemployment Insurance wage records not all workers are included. It is therefore very important to understand what is left out so that if comparisons are made with Bureau of Labor Statistics information from the Current Population Survey, the differences that might arise can be better understood. The LEHD program is supported by several years of research and is a high quality example of the use of administrative data in place of a probability-based survey. Other uses of big data may not examine data quality and data errors, and in that situation using them as an alternative to a probability sample may not be warranted.

A technique that is related to the concept behind big data is *meta-analysis* (Hedges and Olkin 1985). One of the main functions of meta-analysis is to combine data from several studies to improve the understanding of the relationships in the variables. Systematic reviews such as those undertaken by the Campbell Collaboration (http://www.campbellcollaboration.org/) are another

variant of this idea. In the 2012 election, one form of meta-analysis that drew wide attention was *poll aggregation*. The aggregators did quite well in terms of predicting the winner of the election and the percentage difference between the candidates. Aggregators typically combine a series of polls conducted by different survey organizations to attempt to reduce the variance of the estimate.

The general concept of poll aggregation (and meta-analysis) is that larger sample sizes (what the aggregate of polls give you) reduce the variance of the estimates. This is an extremely well known principle that we know works in most cases. For example, if we want a more precise estimate from a survey, increasing the sample size will help. Of course, if the survey has a bias due to measurement, nonresponse or coverage error, then the accuracy of the estimate will not improve as the sample size increases because the bias is not a function of the sample size (and the bias will often swamp the variance).

In the recent U.S. election, there were several poll aggregators but they had different accuracy even though they all probably had access to the same polls. Why? We suspect the answer is that the rules for aggregating differed, with some choosing to exclude polls that used certain surveying techniques, were done too long ago, had poor track records, or were outliers compared to the other polls. Essentially, the aggregators were making a choice to include only polls that they viewed as surveying the same population so they could take advantage of aggregating. If they chose poorly, even a few aberrant polls could bias the aggregates.

The same lesson applies to online research that claims to reduce biases in estimates from panels by aggregating or using different panels. This approach only works when the panels that are aggregated are actually measuring the same population in the same way. The ability to choose which panels should be aggregated is limited unless there is some track record similar to what is used by election aggregators.

As of this writing big data (not the meta-analysis variant) remains more conceptual than practical, but we can expect that to change rapidly and with the likelihood that at least some data currently

collected by surveys will come from these sources. However, we suspect many of the same issues that non-probability sampling faces in terms of incomplete coverage of the entire population as well as different forms of measurement error will be pertinent to big data.

3.6 Summary

Our goal in this section has been to set out two fundamental points in advance of the detailed discussions that follow. The first is that, unlike probability sample, non-probability sampling comes in many forms, each with a somewhat different set of assumptions and procedures for creating samples from which we can make inferences to a larger population. The second is that, despite this variety, there is a shared set of obstacles for the methods to overcome in order to demonstrate the validity of their estimates. Those obstacles include the exclusion of large numbers of people from the sampling process; the frequent reliance on volunteers with few controls; and generally high levels of nonresponse, although we have this last problem with probability samples as well.

Commentators sometimes express these obstacles in different ways and it may be useful to consider a framework suggested by Kish. Kish (1987) described four different classes of variables (p. 2-3):

- *Explanatory variables* that embody the aims of the research design and are the independent and dependent variables among which we expect to find a relationship. These are the items we collect in the survey questionnaire.
- *Controlled variables* are extraneous variables that can be adequately controlled either at the time of selection or at estimation. Geographic or demographic variables used either in sample selection or in post stratification are common examples.
- *Disturbing variables* are uncontrolled extraneous variables that may be confounded with the explanatory variables. These are unmeasured covariates with the measures of interest.
- *Randomized* variables are uncontrolled extraneous variables that are treated as random errors.

The challenge for non-probability methods is to identify any disturbing variables and bring them under control, in sample selection, estimation, or both.

In an experimental context, randomizing the observations after conditioning on the controlled variables yields a randomized experiment that can be used to make valid estimates of causal effects because the randomization ensures than any disturbing variables have, on average, the same effect on

both the control and experimental groups. Similarly, a probability sample mitigates the effect of disturbing variables that might cause self-selected non-probability samples to produce biased estimates for the target population. High rates of nonresponse work against this inherent advantage of probability sampling, but this is sometimes manageable when sample frames contain rich data about sampled members or repeated contacts provide at least some information about the differences between those who respond and those who do not.

Non-probability methods have no such advantages. The selection bias in most non-probability methods creates the substantial risk that the distribution of important covariates in the sample will differ significantly from their distribution in the target population and to such an extent that inferences are misleading if not simply wrong. To be of value non-probability samples must rely on some form of statistical adjustment to manage this risk of large biases. The effectiveness of those adjustments depends on the identification of important covariates, their availability and quality. The integrity of any non-probability method depends on how well it solves this fundamental problem.

4. SAMPLE MATCHING

Sample matching is an approach to non-probability sampling that has been used for many years and across a variety of subject matter areas. Its primary objective in comparative studies is to reduce bias in computing estimates of differences between two alternatives (treatments or interventions) by matching a sample to a control group using one or more characteristics. The characteristics used -- referred to as covariates in the following discussion -- are thought to be related in important ways to the explanatory variables and the outcomes. In non-probability samples used to make inferences to a larger population the goal is to match the sample to be surveyed with the population so that sample estimates are more representative than could be achieved without such a control. The main problem is that uncontrolled covariates (what Kish (1987) calls disturbing variables) can increase bias. Sample matching attempts to overcome this. The basic techniques are well established in evaluation research and the analysis of observational studies. In recent years, sample matching has been applied in more general settings including market, public opinion, and other social science research.

Sample matching is an intuitive and appealing approach to reduce selection biases in non-probability samples. Quota sampling is one well-known and widely used method of sample matching. In quota sampling the objective is to obtain interviews from a specific number of subjects that match the population, typically in terms of readily available characteristics such as age and gender. The appeal of this technique is that the 'matched' sample mirrors the target population in terms of these characteristics and, presumably, bias is reduced

Much of the theory of matched sampling was developed for observational studies. In that setting, a non-probability sample is selected by matching the characteristics of "treated" units to a set of "control" units with the goal of estimating the causal effect of the treatment. For example, a treatment might be an advertising campaign to reduce smoking in a city and the goal of the study is to estimate the effectiveness of the treatment in actually reducing smoking. Effectiveness is measured by comparing data from a city with the treatment to a matched city without the treatment. The

matching is assumed to reduce selection bias as long as the matched or control city has the same distribution of important covariates as the treated city.

We note that while most observational studies use non-probability samples, probability samples also are sometimes used to assess causal hypotheses. Although probability samples may have the advantage of being representative of the target population, that alone does not make them suitable for analysis of causal relationships. For example, a probability sample of adults that captures data on lifetime smoking and cancer is not sufficient to show a causal relationship between the two groups because important covariates may not be distributed equally between the two groups without explicit randomization of those groups.

Thus, sample matching attempts to reduce selection bias much like weighting (see Section 6), that is, by taking advantage of auxiliary or covariate data. Matching uses those auxiliary data in sample selection while weighting does so after the fact. Rubin (1979) recommended using both sample matching and weighting for observational studies of causal effects. Today, similar approaches are becoming popular beyond studies of causal effects.

Matching can be done in a variety of ways. For example, matching can be done at the individual level (like our city example) where for each case or treated unit one or more controls are found that matches it. Frequency matching is still another approach; with frequency matching the goal is to match the distribution of the characteristics in the control sample with the distribution in the treated sample. Most quota samples use frequency matching.

In theory, selection bias is reduced if the characteristics used for matching are the primary variables related to the outcomes and so the covariates are balanced, that is, their distributions in the sample are the same as their distributions in the target population. Rosenbaum and Rubin (1983) describe balance in detail. One of the distinct advantages of random sampling is that it produces covariate balance automatically, even though it does not result in perfect matches. In non-probability sampling, the important confounders may be unknown or not available so that covariates are not

balanced. Covariate balance is essential for valid inference and therefore the ability of sample matching to reduce bias depends on the identification, availability and quality of the auxiliary variables assumed to be covariates.

In this section, we describe some relevant ideas and practices in evaluation research and then discuss newer applications of sample matching for general population surveys.

4.1 Randomized Controlled Trials

Evaluation research almost always involves the collection of information through surveys and/or from administrative records. The typical goal of evaluation research is to establish a causal relationship between a treatment and an outcome (e.g., a specific type of change in benefits will cause an increase in food security). The gold standard of evaluation research is randomized controlled trials, which are characterized by random assignment of subjects to a treatment group (or to one of multiple treatment groups) or to a control group (Shadish et al. 2001). The subjects in the treatment group receive the intervention while the control group subjects do not.

We use the term "gold standard" with some qualification. Although randomized controlled trials are strong on *internal validity* (sometimes called "truth within study"), they may not be strong with respect to *external validity* (sometimes called "truth beyond study"). In theory, measures of the effect of the intervention for the treatment group are not due to confounding factors because randomization automatically creates covariate balance. And so the observed changes in the treatment group reasonably can be attributed to the intervention and not to other causes or factors (confounders). However, the question remains whether the results from the trial extend beyond the particular sample.

Using the example of social program evaluation, we might first identify a group of sites that are willing to participate in an evaluation. Each site provides a list of households that meet the eligibility criteria for the program. The researcher divides the list of households randomly into two groups – one half of the households are designated to receive the treatment and the other half not. The use of

random assignment promotes internal validity. That is, under ideal circumstances the difference observed between the treatment and control groups is due to the program (i.e., the program effect).

If, however, we are evaluating a national program, there are potential limitations that are rooted in sampling. Site selection is a key issue. Was the experiment carried out in only one site, in multiple sites, and how were the sites selected? In many program evaluations, sites apply to participate in the demonstration and the funder selects a judgmental sample of sites to participate. So site selection criteria are far from random. Past participation in other program evaluations, active versus passive consent for receipt of the treatment, and quality and completeness of the list of eligible households may be taken into account in the selection of the sites. Thus, the survey research concept of generalizability or external validity is compromised because potentially important covariates in site selection are uncontrolled. The study might well be valid within the sites selected, but the ability to make inferences to the full population is compromised.

Replication of the experiment in different areas of the country with different populations is one way to reduce the threat to external validity. A larger number of sites may increase external validity but any time a non-probability method of selecting the sites is used external validity remains a concern. Constructing a list of all eligible sites, stratifying the list of key site level variables, and drawing a probability sample of sites is rarely done because it simply is too difficult and expensive to implement in practice.

4.2 Quasi-Experimental Designs

Although randomized experiments may be the gold standard in evaluation research, significant ethical, legal, legislative, and other considerations may preclude the use of randomization in certain types of studies (e.g., denial of treatment in clinical trials). Where these concerns exist quasi-experimental designs are often used. These designs do not randomize eligible sample units (e.g., households or persons) to a treatment group; they rely on other techniques to ensure internal validity to the greatest extent possible (Shadish et al. 2001).

Sample matching is probably the most popular of these techniques. Examples include designating sites to receive the treatment and then identifying similar sites to be part of the comparison group. Other designs involve designating households or persons within a site to participate in the demonstration and then identifying a matched control sample for the comparison group in that site. The ability to match, for example, at the household level is a function of the variables that are available in the administrative data at the site level.

A related technique that has received considerable attention in evaluation research is *regression discontinuity design*. The basic idea is that sites that just fail to meet the eligibility criteria (e.g., the percent of children in the school district that receive free or reduced price school lunches is just above some threshold for eligibility) are taken as the treatment group. Sites that are not eligible because they are just above the program eligibility threshold are designated as the comparison group. This, too, is a form of sample matching. One potential threat to external validity arises when the treatment and comparison sites are compressed around the eligibility threshold and the treatment effect measured in the study may not generalize well to sites that are distant from that threshold.

Sample matching can be especially challenging when there are a large number of covariates. One solution frequently discussed in the evaluation research literature is the use of *propensity score modeling*. The goal of sample matching is to ensure that the distribution of observed covariates is similar across the treated and control groups. When used in this context, the propensity score is the conditional probability of receiving the treatment rather than the control given the observed covariates (Rosenbaum and Rubin 1983). Since the propensity score is unknown it is usually estimated from the available auxiliary data using logistic regression. Matching is now simpler since only one variable, the propensity score, is used, although in some cases matching is done using the propensity score and another important covariate. Rosenbaum and Rubin (1983) argue that matching on propensity score alone does the same in terms of removing selection bias as matching on the full set of covariates.

Some have suggested that some quasi-experimental designs that use propensity score adjustments are superior to randomized designs but the evidence is not strong (Rubin 2008).

4.3 Medical Studies and Clinical Trials

Medical and clinical trials use a broad set of techniques beyond the randomized controlled study and the quasi-experimental studies described above. Many of these studies rely on volunteers. They assume that the effect of a new medical procedure observed among the volunteers is likely to hold in the general population. This is a very strong assumption that may not be valid.

Types of studies include:

- *Randomized Controlled Clinical Trials*. Recruited subjects are randomly assigned to two or more treatment groups at the start of the study and the outcomes observed in the groups are typically compared at two or more points in time.
- *Randomized Cross-Over Clinical Trials*. Subjects with a disease/medical condition are randomly assigned to one of two treatment groups at the start of the study and after a time period appropriate for the treatments the subjects are switched to the other treatment for the same period of time.
- *Randomized Controlled Laboratory Study*. This refers to experimental studies using animals with data collected in a laboratory environment.
- *Cohort (Incidence, Longitudinal Study) Studies*. A group of subjects is identified such that some of them have had or will have the relevant exposure; the subjects are studied at one or more follow-up point(s) in time to estimate the strength of the association between the exposure being studied and one or more outcomes.

As we have seen, most medical studies and clinical trials are based on non-probability sampling techniques and stress internal validity. No study is useful for external validity if it does not have internal validity. However, clinical trials guidelines give a great deal of attention to statistical power and sample size determination, but little or no time is spent on generalizability. As a result, many of these studies do not hold up when attempts to replicate the findings are mounted (Mayes, Horwitz, and Feinstein 1988; Young and Karr 2011).

4.4 Sample Matching for Surveys

The goals of the types of studies described thus far are to make inferences about the effect of a treatment or intervention. In other words, they try to address questions about whether the treatment

caused the outcome to change and, if so, by how much and in what direction. More recently, the ideas proposed in this literature have been used to draw samples for surveys that are more descriptive and not aimed at understanding a specific causal relationship. Thus, the application of sample matching is somewhat different in the survey context.

The thrust of most of the research in sample matching methods for surveys has been to match background characteristics of the selected sample to the target population. In this setting, the target population, say the U.S. household population is similar to the treated group in the evaluation literature (or to the cases in the case/control designs). The matched sample is the equivalent of the control group. Just as with other forms of research discussed above, the idea is that even with non-random samples researchers will be able to make inferences to the target population because the sample matching will balance the covariates so that selection bias is reduced and the survey estimates will mirror those of the population.

In market research the target population might be the household population as described by the national statistical office. In election research, the target population might be the population of likely voters. In these settings, sample matching is intended to remove or reduce the selection bias so that sample estimates (e.g., percent voting for a specific candidate) are more likely to be accurate estimates of the population parameters (e.g., percent of election-day votes cast for a specific candidate). The estimates are made only from the matched sample in this situation. To the extent that selection bias is not removed we can expect estimates also to be biased because an important covariate was not accounted for in matching.

The differences between sample matching in causal analysis versus surveys are important and worth discussing in more detail. First, in causal analysis the inferential objective is clear and the sample matching is focused on identifying matching variables that are related to the outcome or treatment effects (e.g., age, sex, genetic predisposition, and body mass index might be considered as important matching variables for studying the effect of a diet to reduce heart attacks because these

variables are related to the outcome). For surveys, typically many outcomes (estimates) are of interest and these may have very different covariates that would be used for matching than if the goal of the survey were only to produce a single estimate. This difference makes the choice of matching variables for survey applications more difficult.

A second difference is that survey estimates are typically expected to generalize to a large population such as all adults or likely voters. One of the techniques used in many observational and medical studies to increase the internal validity of the results is to restrict the population to a very specific group (women between the ages of 18 and 40 who have never had children or adults who stopped smoking after having smoked for more than 10 years). Restricting the study to the subgroup helps ensure that subgroups are as similar as possible, and so selection biases are reduced. (Rosenbaum 2005). In most public opinion and market research surveys, this approach is not appropriate because estimates of the entire population or a large and very diverse subpopulation are required. Point estimates (e.g., the proportion supporting a particular policy or candidate) are also important in surveys whereas in clinical trials the focus often is on whether treatment is more effective than the control (i.e., the difference is sufficiently large) rather than on the exact proportion who will benefit from a treatment.

We currently are aware of no standard sample matching practices for collecting non-random, non-experimental data that generally support inferences to a larger population. This is in contrast to the probability sampling paradigm. The general approach to matching is to identify a constellation of background variables that that might be disturbing or confounding variables. As in causal studies, these are characteristics that tend to co-vary with the estimates produced from the survey and may be different in the non-random sample when compared to the target population.

Some telephone studies (especially in market research) and much of the online research typically employ sample matching methods to compensate for nonrandom sampling. For example, telephone studies may include age, gender, and geography quota cells that specify the number of observations

required in each cell. The specified quota cells represent a model, a proxy, for what the researcher would expect to obtain if a valid simple random sample could be drawn. This may be done because of nonresponse rather than sampling issues; for example, uncontrolled telephone surveys typically obtain a higher percentage of female respondents than males compared to their relative population percentages. Studies based on opt-in panels often include quota controls for the same reason. These are all non-random research designs where the researcher attempts to build a representative dataset using basic sample matching techniques.

Some survey researchers have begun using more complex methods of sample matching similar to the methods used in observational studies (Rosenbaum 2005). These methods generally rely on a larger and more varied set of covariates than those used in causal studies. We describe the methods in the non-probability survey context below while noting that these applications have not been documented as extensively as those for causal studies.

4.5 Current Uses of Sample Matching in Surveys

The process typically begins by identifying the target population that the survey will make inferences about, say all adults in the U.S. or likely voters. Characteristics of the population are then obtained (estimated) from some source or combination of sources deemed of high quality such as the American Community Survey (ACS), the Current Population Survey (CPS), the General Social Survey (GSS), the American National Election Survey (ANES), or from a probability-based survey designed specifically to support the sample matching effort. The next step varies depending on the survey organization's approach.

Vavreck and Rivers (2008) described one approach using the principles outlined in Rivers (2007). They first selected a random sample of 38,000 persons from the publicly released ACS file and used this as the cases in traditional case control studies. Next, they found the closest match in the pool of available persons from within their opt-in panel of volunteers for each unit in the ACS. They used a distance function defined to ensure similarity across the characteristics used in matching (known for

both the ACS and the panel respondents). Four observed variables (age, race, gender, and education) plus imputed variables measuring partisanship and ideology were defined as covariates and used in the matching. The matched sample from the panel was then invited to participate in the survey and their responses were used to produce the survey estimates after some additional statistical weighting procedures to account for nonresponse.

This approach is an example of a *one-to-one matching* procedure, where the goal is to have one respondent who matches the 'case' or in this situation the targeted U.S. adult population as sampled from the ACS. In causal analysis, the estimation of causal effects compares the cases and the controls (Rubin 2008). In the survey setting, causal effects are not being estimated and only the data from the matched sample (the opt-in panel in Vavreck and Rivers) is used to produce the survey estimates. As a result, sample matching for surveys does not maintain one of the useful attributes present in causal studies where the matching introduces a correlation between cases and controls to improve the efficiency and the robustness of the estimated treatment effect (Rubin and Thomas 1996).

When many auxiliary variables are needed and available to reduce selection biases, one-to-one matching can be problematic. It may require a very large panel to match exactly on each of the characteristics because the number of matching cells increases geometrically with the number of matching variables and levels. Vavreck and Rivers (2008) dealt with this problem by limiting the number of matching variables and using a distance measure rather than trying to fill cells.

As we described above, a standard approach in observational studies when there are a large number of covariates is to match on propensity scores rather than specific characteristics. This approach transforms the multi-dimensional problem with a potentially huge number of cells into a more easily-managed univariate problem. Since the propensity score is a continuous variable, the score may be categorized and then matches are found within the same category or a distance measure may be used to find the closest match. While propensity scores have been used in the statistical

adjustment phases of non-probability samples (e.g., Terhanian et al. 2001; Lee and Valliant 2009), they have only recently been introduced for matched sampling in non-probability surveys (Rivers 2007, Terhanian and Bremer 2012).

Frequency matching is another approach to sample matching that is more commonly used in observational studies than one-to-one matching. With frequency matching, the frequency distribution of the target population is estimated (the percentage distribution of persons by age, sex, etc.) and the matched sample is chosen to ensure that the frequencies of the matching variables are similar to those of the target population. There are a variety of methods for frequency matching including category matching, caliper matching, stratified random sampling, or a variant of pair matching (e.g., Rothman and Greenland 1998).

Eggers and Drake (2011) have described a version of frequency matching that they call *dynamic quota cell-based sample matching*. With this method, the researcher begins with the GSS as the source of the target distributions. They prefer the GSS because it contains questions for multiple dimensions such as demography, psychology, and behavior. The researcher draws a matched sample from an opt-in panel to frequency match to the GSS variable distributions (the method does not require a sample from the GSS). Since some matching variables of interest may not be available on the panel, they recommend a small set of GSS questions be asked of panel members to finalize the matched sample. The final matched sample is sent the survey and their responses weighted and tabulated to produce the estimates

Terhanian and Bremer (2012) have described another approach. They traced the origins of their approach back to Cochran, Tukey and Mosteller (1954) and the concept of parallel surveys. As originally implemented by Harris Interactive, the method used propensity score models to attempt to reduce the bias in a non-probability sample at the weighting stage by aligning the distributions of covariates in the sample with those of a probability-based reference survey of the target population (Terhanian et al. 2001; Terhanian 2008). A key potential weakness of this approach is that it relies on

post hoc adjustments that can be especially problematic if the biases in the original sample are large. Under such circumstances weights can become very large, reducing the effective sample size precision of estimates.

Terhanian and Bremer (2012) have attempted to solve these and other related problems by adapting matching in sample selection, and retaining the weighting procedures. Their *Propensity Score Select* methodology also relies on parallel surveys (an RDD telephone survey and a survey using an opt-in panel). Both surveys should have the same target population and use the same questionnaire or at least a shared subset of items. In addition to the substantive questions from which the measures of interest are to be estimated, both surveys should have questions that ideally "account for all observed and unobserved differences" between the target population and the opt-in panel. These covariates may be demographic, attitudinal or behavioral. The researcher also may want to include some benchmark questions that can later be used to test the external validity of the estimates. Post data collection covariates are used as independent variables in a logistic regression model to reduce the differences across the two surveys on the items chosen to improve external validity. The model can then be used as the basis for sample selection in future studies with the same target population on the same or similar topics. The method seems best suited to very large data collections where a pilot can be run to develop the model or for tracking research.

Finally, Gittelman and Trimarchi (2009) have described an approach they call *the grand mean*. Their primary goal was to maintain consistency and reliability from sample to sample, rather than improve external validity. They administered the same questionnaire to samples of 500 respondents each, drawn from over 200 panels in 35 different countries. From these data they developed a segmentation scheme that classifies people by their buying behavior, media consumption and sociographics in seven different market segments (automotive, consumer electronics, banking, etc.). These segments can be used to type prospective respondents from almost any sample source and samples drawn to match the proper distributions within each segment crossed by the market segment

being studied. More recently, they have begun to include an RDD portion in the samples used to develop their segments (Gittelman and Trimarchi 2010). By doing so they believe that they increase the representativeness of panel samples. They continue to develop this approach and a needed part of that development is the explicit identification of their assumptions.

4.6 Summary

The matched sampling techniques discussed above are interesting and innovative applications for surveys that rely on methods used in other fields for many years. While our discussion of those methods generally has shied away from reporting on the validation studies that demonstrate their effectiveness, such studies exist and readers are encouraged to examine them before judging the validity of the method. But we also note that there are some key assumptions running through all of these methods and consumers of their results should consider them carefully.

First, the data used for the control group must be high quality and have minimal error. All three of the methods we described rely on survey data and assessing the degree of error in the reference survey(s) is important.

Second, we need to bear in mind that these techniques were originally developed to estimate a single treatment effect while surveys are used to produce many estimates. In causal studies treatment effects are estimated using data from both the treated cases and the matched sample (the estimated effects being the difference between these two groups). In the survey setting, the data for the matched sample alone is used to produce estimates. The robustness and reliability of the inference when only the matched sample is used to produce the estimates has not been fully examined.

Third, and perhaps most important, surveys typically cover a large number of topics and even a single survey is often used to generate many estimates. Different topics have different covariates. No single set of covariates can be expected to correct all the bias in a full range of survey topics and the number of covariates needed for a given survey may be quite large. As we have said before in this report, a key assumption of any non-probability method is that the key confounders or covariates

have been identified, measured, and balanced. If this has not been done well then sampling will result in serious selection biases.

As Rosenbaum (2005) noted, "Because even the most carefully designed observational study will have weaknesses and ambiguities, a single observational study is often not decisive, and replication is often necessary. In replicating an observational study, one should seek to replicate the actual treatment effects, if any, without replicating any biases that may have affected the original study." For surveys, where the negative consequences of small errors in the estimates may be much less severe than in estimating a treatment effect, this sage advice may not be practical. However, it is feasible for those using matched sampling for surveys to consider methods for replication that are practical. For example, if several matched sample surveys are being conducted, some common items could be included so that replication could be achieved without adding greatly to costs.

In other fields that rely on matched studies, researchers are aware and concerned that the results have weaknesses that sometimes lead to conclusions that are later contradicted. For example, Mayes, Horwitz, and Feinstein (1988) described 56 topics investigated in case-control studies that were later contradicted by other research. They noted, "Contradictions can arise whenever causal relationships are investigated in studies where the compared agents did not receive randomized experimental assignment, and where the groups and data are collected without deliberate strategies to avoid or reduce bias." Non-probability samples using matching methods face the same issues.

Finally, an important issue that we have not covered is transparency in the methods. Transparency is a significant issue in observational studies and is equally important in matched sampling surveys. Vandenbroucke et al. (2007) reports on the activities of a group of methodologists, researchers, and editors who provided recommendations to improve the quality of reporting of observational studies. They call these recommendations the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) recommendations. They provide a valuable

blueprint that could be used as a model for reporting requirements for non-probability sampling in general and most specifically for matched sampling for surveys.

5. NETWORK SAMPLING

Suppose you are studying the population of men who have sex with men (MSM) in a small city. You expect that MSM constitute about 2% of the male population of the city. How would you go about sampling them? You could use a standard sampling frame for the full population, then screen each potential respondent to determine if they were MSM before administering the full survey. However, you might need to contact and screen 50 times the number of participants you were seeking. The screening process would also need to be extensive enough to build enough trust for respondents to reveal their MSM status. This approach would be quite costly and, in many cases, prohibitively so.

Another approach would be to identify and sample MSM from known sites or sources where they congregate. Interviewers could visit known local MSM hangouts, such as bars, parks, or community events, and survey MSM found there. These samples, however, would likely not reach many subgroups of MSM in the city.

Network sampling offers an alternative. Because MSM are often a socially-connected sub-population, their social network connections can be leveraged to facilitate sampling. If a small number of MSM can be identified, and sufficient trust can be established, these first MSM can connect researchers to their social contacts, who can, in turn, connect researchers to further MSM, and so on until the desired sample size is achieved. This approach has proven valuable in attaining large and diverse samples in many so-called hard-to-reach populations, in which traditional survey sampling is infeasible. Inference from such samples is complex, however, as most are not probability samples with known probabilities of selection.

The above is an example of *link-tracing network sampling*, a strategy leveraging network connections to facilitate sampling. The defining feature of this approach is that subsequent sample members are selected from among the network contacts of previous sample members. In this way, the network "links" are traced to enlarge the sample, leveraging the network structure to facilitate recruitment. With the recent rapid expansion of online social networks such as Facebook, network sampling has

become feasible on a very large scale, and appealing as a practical and inexpensive way to reach large numbers of people.

Network sampling is not fundamentally a non-probability sampling approach. In fact, its early foundations in the statistical literature (e.g. Goodman 1961; Frank 1971; Thompson 1992) are probability sampling methods. In more recent practice, however, link-tracing network sampling has proved useful in cases where the strict assumptions required for probability sampling are not applicable. Examples include rare ethnic minorities (e.g. Welsh 1975; Snow et al. 1981), those at risk for disease such as HIV (e.g. Klovdahl et al. 1994), and marginalized workers (Bernhardt et al. 2009).

5.1 Motivation

Link-tracing network samples are often of particular use for hard-to-reach human populations, in which traditional sampling methods are unavailable or impractical. They are often used for cost-savings in cases where traditional methods are available but costly. Although network sampling may reduce the cost of data collection, far more samples may be needed for a given level of precision because the data typically are dependent. Hence, it is not always clear when the cost-per-information of such a strategy is actually lower than that of more traditional approaches. In some populations, standard methods may not even be feasible. For example, sexual minorities may be stigmatized in the larger population and so the identification of target population members in available sampling frames can be difficult (e.g. Zea 2010). Other populations, such as small ethnic minorities, are rare enough that samples from available frames would capture them only at a very slow and inefficient rate (e.g. Kogan et al. 2011).

Kish (1965 and 1987) and Kalton (1993, 2003 and 2009) discuss various probability sampling techniques for rare subpopulations (domains) that allow for valid design-based estimates of characteristics of the rare subpopulation of interest. Probability sampling techniques for rare populations include:

- Building lists (sampling frames) of the rare subpopulation,

- Multiplicity sampling,
- Disproportionate stratified sampling,
- Non-overlapping multiple frame designs,
- Overlapping multiple frame designs,
- Accumulating eligible sample elements from previous cross-sectional surveys,
- Sample designs involving screening for the rare subpopulation,
- Two-phase sampling.

These sampling techniques have been in use for many years. The CDC's National Immunization Survey (Centers for Disease Control and Prevention, 2005) is an example of a screening survey for households containing children age 19 to 35 months. The 2011 Survey of Muslim-Americans (Pew Research Center, 2011) provides a recent example of the use of a multiple frame sample. Only about 0.5% of random digit dial (RDD) respondents self-identify as Muslim or Muslim-Americans, so this sampling strategy combined landline and cell RDD targeted to high-Muslim density areas, with re-contacting previous Muslim RDD respondents from other studies.

However, in some situations it is not practical to use probability sampling for rare subpopulations because techniques such as screening for members of the rare subpopulation is too costly given the desired target number of interviews. Costs can be particularly prohibitive when the rare subpopulation is located in 1% or less of the households in the geographic area being covered.

In other situations probability sampling techniques cannot be applied given the definition of the rare subpopulation. For example, probability sampling techniques oriented to the household sampling frame are not appropriate if a substantial portion of the rare subpopulation is located in group shelters or is homeless. It may not be possible to build a list of a rare population. This is likely to occur when the rare subpopulation is a "hidden population," for example, those participating in an illicit behavior such as using intravenous drugs.

Other non-probability methods for sampling from such populations are available. Three prominent alternatives are quota sampling, targeted sampling, and time-location sampling. Quota and targeted sampling are both purely non-probability methods.

Targeted sampling (Watters and Biernacki 1989) is a non-probability sampling method that combines extensive ethnographic mapping with sampling quotas, time and location quotas, and peer-referrals constituting network sampling. This is a careful, pragmatic, though non-probabilistic approach, designed to gather a representative sample of a hard-to-reach population. It has been used with some success (Carlson et al. 1994). Robinson et al. (2006) conducted a comparison of targeted sampling and respondent-driven sampling (RDS), a network-sampling variant discussed in detail below. They found comparable quality of samples resulting from the two methods, but with greater staff effort required for targeted sampling, and greater financial incentives required for respondents in respondent-driven sampling.

Time-location sampling (Muhib et al. 2001, MacKellar et al. 2007) can be based on a probability sample drawn from a known frame, but from a sampling frame that has two levels. Because the relationship between these frames and the target population is unclear, it is ultimately a non-probability sampling method. The upper level sampling frame is based on times and locations where members of the target population are known to congregate. Such time-locations are treated as strata. Within each stratum, population members are sampled based on a method chosen by the researchers, often a census or a probability sampling strategy. Such an approach is technically a probability sampling strategy, because the observation probabilities of all participants can be determined by their memberships in the defined strata. In practice, the method is severely limited in that it is difficult to construct strata that cover the full target population. In the case of MSM, for example, the targeted times and locations may well cover most centers for gay culture, but might fail to cover segments of the MSM population who do not frequent such establishments. Time-location sampling is very similar to *location-intercept sampling*, in which strategically-placed interviewers systematically recruit passers-by. It can also be considered a form of cluster sampling, where the clusters are designated by

the times and locations under study. As in cluster sampling, it is often possible to draw a probability sample from within each cluster, but the relationship between the clusters and the larger target population is unclear.

An advantage of network sampling is that the social ties of respondents are able to extend the sampling frame beyond the visible or accessible population members, hopefully in a manner that allows the sample to be treated as a probability sample.

In Brazil, for example, Kendall et al. (2008) found the RDS network-sampling strategy yielded less-expensive, more diverse samples of men who have sex with men than time-location sampling. This is often true for populations that are rare, stigmatized, or living outside of households. Because of the dependencies between successive samples, network samples are often not the first choice of surveyors. In many cases, however, they represent the most principled viable alternative.

5.2 Statistical History

Although introduced earlier, pre-dating even Coleman (Coleman 1953; see also Handcock and Gile 2011), link-tracing sampling is often traced back to Goodman's 1961 formulation of snowball sampling. In this work, Goodman introduced a variant of link-tracing network sampling which he refers to as *s stage k name snowball sampling*. This original probabilistic formulation assumes a complete sampling frame is available, and the initial probability sample, or *seeds*, is drawn from this frame. *k* contacts of each seed are enrolled in the first stage, or wave of the snowball. *k* contacts of each wave-one respondent form the second wave, and so forth, until s waves are sampled. Goodman advocates this sampling strategy for inference concerning the number of network structures of various types in the full network. In particular, he focused on cycles in the network: polygons such as triangles, squares, etc. In contrast to the shortcomings of link-tracing samples as currently practiced, he advocates snowball sampling as a method of increasing the efficiency of the sample. In particular, he argues that by using a s-stage snowball to estimate the number of s-polygons (for example, 3-waves for triangles, 4-waves for squares), far fewer nodes need be observed for the

desired precision in the estimated number of s-polygons than would be required based on a simple random sample of nodes.

Ove Frank's work, beginning in the 1970's (e.g. Frank 1971, summary in Frank 2005), built on Goodman's formulation, and systematically expanded the types of samples, objects of inference, and estimators for which link-tracing samples could be used. Like Goodman's work, Frank treated cases in which the seeds could be drawn using a probability sample from the target population, and in which all network features necessary for enumerating sampling probabilities could be observed. Known sampling probabilities are necessary for inference in a design-based framework. Handcock and Gile (2011) discussed the limitations these conditions impose for link-tracing network samples and design-based inference, and concluded that many of these conditions are unlikely to be reasonable in practical situations.

Multiplicity sampling (introduced by Birnbaum and Sirken 1965) is a special class of link-tracing network sampling designs in which sampling probabilities are available. In such samples, a sampling frame is constructed in such a way that some members may be sampled more than once. For example, in Brick (1990), youth were over-sampled by conducting a household telephone survey and asking about any youth, either living in the home or children of women living in the home. This approach has also been used to over-sample those with rare diseases, or certain ethnic minorities (see Kalton and Anderson 1986 for a review). Multiplicity sampling is based on following only one step, or wave of network sampling, according to well-defined relations such as kinship ties or geographical adjacency, allowing for known ratios of sampling probabilities.

The work of Steve Thompson and co-authors (e.g. Thompson 1990, 2006a, 2006b; Felix-Medina and Thompson 2004) moved multi-wave probability samples from link-tracing designs to the realm of realistic sampling and inference. In particular, Thompson (1992) and Thompson and Seber (1996) highlighted the adaptive nature of many link-tracing designs. In a conventional sample, the sampling design is fully specified prior to data collection, including the sampling probabilities of all units at each stage. In an adaptive design, information collected during the study can influence future data

collection. Network samples are often of this type: network ties to previous respondents discovered during sampling can influence sampling probabilities. For valid statistical inference with available methods, however, it is important that the sampling process depend only on features of the network that are observed. Remarkably, link-tracing network samples begun with probability samples often satisfy this criterion.

Thompson and colleagues (Thompson and Frank 2000; Thompson 2006a,b) proposed many methods for sampling and inference using adaptive sampling strategies. The limitation of most of these methods is that they require an initial probability sample, and hence a sampling frame of the target population (the exception is the works for Felix-Medina and colleagues, beginning with Felix-Medina and Thompson 2004, which requires a partial frame of venues). In many practical settings such a frame is not available and this is, in fact, the motivation for the network-sampling strategy.

Of special note is the work of Thompson and Frank (2000), which introduced a mode for likelihood inference from link-tracing samples, as opposed to the earlier design-based framework. Statistical models can be fit to link-tracing network samples, without modeling the sampling process and without necessarily knowing the sampling probabilities of each unit in the sample whenever the sampling depends only on the observed part of the network. Again, however, this approach does require an initial probability sample.

5.3 Non-probability Network Samples

In many contexts, the phrase "snowball sampling" has come to mean something very different from the above probabilistic formulations. It sometimes means a convenience sample (see Section 3) acquired by starting with a non-probability sample, then expanding the sample by enrolling (typically all) of the contacts of each previous participant. Such samples are clearly not probability samples, as the probability of being sampled is determined first by the initial sample of convenience, and subsequently by having network connections to the earlier convenience sample (see, e.g. Biernacki and Waldorf 1981; Handcock and Gile 2011).

Examples of the use of snowball samples in a non-probabilistic formulation are quite varied, and include Trow (1957), who studied political radicals in Bennington, VT, Kaplan et al. (1987), who studied users of heroin in the Netherlands, and McKenzie and Mistiaen (2009), who studied families of Japanese descent in Brazil.

5.4 Online Network Sampling

It is difficult to mention social networks without conjuring images of the Internet, both in general and in specific contexts such as Facebook or Twiter. Indeed, these and other online forums for social connection provide opportunities for sampling. Online social network sampling has been pioneered in the electrical engineering and computer science literatures, where software-based "crawlers" are designed to follow a specified algorithm to trace internet links (such as Facebook friendships) from one user to another, gathering data from the available part of each sampled user's account. The crawling algorithms used are varied and the subject of a great deal of research. Gjoka et al. (2011), for example, compared several naïve approaches, as well as introducing more nuanced crawling to approximate representative samples from the networks. Other specialized algorithms aim to address other challenging network features. *Frontier Sampling*, for example, aims to address the possibility that some parts of the network may not be connected through links to other parts of the network (Ribeiro and Towsley, 2010). Approaches in this vein, however, currently do not ask users to complete new survey questions, but instead collect information available in online content.

The use of online social networks to facilitate sampling in survey research is not well developed. Wejnert and Heckathorn (2007) introduced an Internet-mediated version of RDS (described in detail below), which they refer to as *WebRDS*. Respondents recruited their contacts through e-mail, and surveys were completed online. They implemented such a study among college students, and found this method to be efficient at recruiting large numbers of students in a short time frame. They were able to recover some known features of the undergraduate population with reasonable accuracy, but

some estimates were subject to substantial biases because of the differential connection to e-mail of various sub-populations.

5.5 Respondent-driven Sampling: Approximating a Probability Sample

Respondent-driven Sampling (Heckathorn 1997) is an approach to sampling and inference that attempts to mediate between the practicality of a convenience sample and the inferential capabilities of a probability sample. This seeming miracle of mediation is attempted with two features: clever sampling design and a healthy dose of assumptions, although some are virtually untestable.

RDS sampling design includes two key innovations, both related to recruitment via the passing of coupons. The method draws its name from the respondent-driven nature of coupon-passing. In most link-tracing sampling designs, respondents are asked to list their contacts in the target population and researchers select from among them to complete the sample, a process that can raise confidentiality concerns in stigmatized populations. In RDS, respondents are given uniquely identified coupons to pass to some of their contacts, making them eligible for participation. This greatly reduces the confidentiality concerns, and makes sampling practical in a much wider range of populations.

The second key innovation is in limiting the number of recruits selected by each participant. In contrast, other link-tracing samples often enroll as many contacts of each participant as possible. This innovation allows RDS samples of a given sample size to move more social contact steps away from the initial sample than other link-tracing samples. For comparison, if respondents enroll an average of five contacts in the study, a sample starting with one seed will likely enroll over 150 respondents within three waves. If an average of two contacts of each respondent are recruited, enrolling 150 respondents from a single seed will likely require more than six waves. More waves mean the resulting samples are less dependent on the initial sample. Many inferential methods therefore use this reduced dependence to treat the resulting data as a probability sample.

Clearly, treating the data as a probability sample requires further assumptions. These vary by estimator, and new estimators are being introduced that change the set of assumptions required (typically by substituting one set of assumptions for another). Here, we will describe the assumptions of the estimator introduced by Volz and Heckathorn (2008). Known as the *RDS-II* or *VH* estimator, it is also closely related to more standard statistical methods, particularly those discussed in Hansen and Hurwitz (1943). This is described in more detail in Gile and Handcock (2010).

One set of assumptions is required to remove dependence on the initial sample or seeds. Suppose population members in a target city only have ties to others of the same neighborhood. Then regardless of how many waves one samples, a sample starting in one neighborhood will never cross over to another neighborhood. Thus, the neighborhood composition of the sample would be fully determined by the initial sample. This is a very extreme version of *homophily*, the tendency for people to be tied to other people like themselves more often than at random (McPherson et al. 2001). Even relatively weak forms of homophily can result in RDS samples being unduly influenced by the composition of the initial sample. Multiple waves of sampling sometimes are needed to counteract this tendency. Furthermore, if parts of the network are completely unreachable from other parts, it is impossible for samples starting in one sub-population to reach the other. This is called a *disconnected graph*. A VH estimator assumes the graph is connected, and that homophily is sufficiently weak for the number of waves sampled.

Because this is design-based inference, this approach also requires known sampling probabilities for all observed units. The VH estimator assumes that these probabilities are proportional to each respondent's degree, or number of contacts in the target population. Intuitively, this makes sense because more connected individuals are more likely to be sampled. This is based on modeling the sampling process as a Markov chain on the space of population members (Gile and Handcock 2010 describe this in detail). If the process were truly a Markov chain, each respondent would recruit exactly one social contact, chosen completely at random from among his contacts. Aside from the branching structure, one concern with this approximation is that it assumes people are not prohibited

from re-sampling previous participants. In practice the sample is without-replacement, so a contact that has already participated cannot participate again. Therefore, the Markov chain approximation requires a large population size with respect to the size of the sample. If all ties are reciprocated, that is, Joe is Sue's contact if and only if Sue is Joe's contact, then under this approximation participants' sampling probabilities are proportional to their degrees. Estimating these sampling probabilities also therefore requires accurate measurement of degrees.

Newer estimators have been introduced that relax some of these assumptions. The estimator in Gile (2011) directly treats the without-replacement sampling assumption so that the population size need not be large. Because it treats a potentially finite population, however, it requires an estimate of the size of that population. The estimator in Gile and Handcock (2011) addresses the without-replacement assumption, and introduces an approach to reduce the dependence on the seeds, and relax the requirement of low homophily or sufficient sample waves. The new approach relaxes these assumptions but requires more reliance on information being available in the sample for estimating the level of homophily.

Although these estimators may provide nearly unbiased point estimates under these assumptions, the variance of the estimators is still potentially troubling. Because RDS produces dependent samples, each sample adds considerably less information than it would in a simple random sample. For this reason, the variance of the resulting estimates is substantially higher than that of a simple random sample of equal size. This phenomenon is exacerbated by higher dependence, resulting from higher homophily. Salganik (2006) and Goel and Salganik (2009, 2010) studied this point in greater detail.

The estimation of uncertainty in RDS is also quite challenging. Salganik (2006), Volz and Heckathorn (2008), Gile (2011) and Gile and Handcock (2011) all provided estimators for the standard errors of RDS estimators. While each providef a valuable approximation, none accounted for all potential sources of variability in this complex sampling process. See Section 7 for further discussion of estimation of uncertainty from non-probability samples.

5.6 Conclusions

Network sampling is both challenging and intriguing. On the one hand, a large number of often untestable assumptions may be required for valid inference. Because of the dependence among sampled units, the variance of resulting estimators can be quite high. For these reasons, network sampling may not be the first choice for a sample design.

On the other hand, network sampling offers a viable approach to some sampling problems that are not readily addressed by other methods. When no valid sampling frame for traditional sampling is available, exploiting the social ties between known and unknown members of the target population may be the only viable means to study a population of interest. If the population is socially connected, network sampling may be the most rigorous method available. Small subgroups defined by immigration status, ethnicity, sexual practices, drug use, or employment type have all been reached by network samples. In such cases, researchers can only be as vigilant as possible in devising a strategy for sampling and inference that makes reasonable assumptions and approximations, allowing for the most principled inference possible.

6. ESTIMATION AND WEIGHT ADJUSTMENT METHODS

Researchers draw samples for studies when data collection from the entire universe (i.e., a census) is not feasible. Depending on the goals of the study, this reliance on a sample may mean that some estimation procedure will be needed to go from the responses from sampled units to valid population inferences. In general, study goals can be divided into two categories. The first category includes studies designed to produce statistics only for the sample in hand. Examples include cognitive testing and pilot studies where sample statistics provide a set of metrics to evaluate study procedures; the goal may be to apply the findings to a much larger subsequent sample but the researchers are not trying to make inferences to the target population. As discussed in Section 4 of this report, another purpose might be to assess the internal validity of certain measures. Although both probability and non-probability designs have been used to obtain the sample statistics for these types of studies, most use non- probability samples to reduce data collection costs.

The second category includes studies designed to make inferences from the sample to a target population by producing estimates. This category of studies is the focus of this section.

We start with the concept of *target population*. Lohr (1999) defines target population as "the complete collection of observations" under study as specified in the survey design. Using one or more sampling frames a probability sample links to the target population by the probability of each individual being selected into the sample. Under the design-based inference paradigm these probabilities of selection make it possible for the researcher to calculate estimates for the target population. By contrast, non-probability sample designs often lack a sampling frame so the linkage is not well defined and probabilities of selection are undefined. As a result, explicit design-based sampling weights that are the inverse of the selection probabilities cannot be produced and so this method of inference is not possible with non-probability samples. This has caused some researchers to argue that questions about the statistical properties of the estimates from non-probability samples

cannot be answered (Biemer and Lyberg 2003, Section 9.2). These concerns are valid within the design-based paradigm. However, it does not necessarily follow that all statistical inferences from non-probability samples are impossible. However, because probability and non-probability samples use different procedures for computing statistical properties and making inferences, some organizations classify the estimates generated from non-probability surveys with terms not used for probability samples. For example, the National Agricultural Statistics Service (NASS) (Matthews 2008) calls estimates generated from non-probability samples "indications." Researchers, such as Couper and Bosnjak (2010), refer to non-probability sample estimates as measures of "internal validity" for the sample and not the population.

Whenever we make inferences from a sample to a target population the statistical properties of the estimates assume great importance. The particular properties of interest are the bias (the estimates are unbiased if, on average, they equal the population values); the variance (the variability of the estimates around their average); and the mean square error (a measure of overall accuracy that is equal to the bias squared plus the variance). The behavior of the bias and variance as the sample size increases is especially critical because we want large samples to be as accurate as possible. We briefly review these concepts below.

Researchers have developed a variety of techniques to improve the statistical properties of population estimates. In probability samples, these techniques are well-documented in survey sampling textbooks and journal articles (see, e.g., Särndal, Swensson, and Wretman 1992). For non-probability samples, the results are more dispersed across disciplines, in part because there are various methods of non-probability sampling. Each of these methods has its own literature and researchers in one field sometimes borrow methods from other fields to address the problems posed when making inferences from non-probability survey samples.

Some of these estimation or adjustment methods are discussed in other sections, such as those on sample matching and network sampling. In this section we provide a somewhat broader review of estimation methodology including analysis weights and procedures that may be used with a variety of

non-probability sampling methods. We avoid most of the technical details but provide references for those wishing to explore the issues more completely.

6.1 Statistical Properties of Estimates

Samples provide a practical method for estimating population values but those estimates generally are not perfect replicas of those population values. These imperfections are rooted in sampling errors (due to not observing the full population) and nonsampling errors (due to inadequacies in measuring the units). Even though these errors are of concern in both probability and non-probability samples, non-probability samples often receive closer scrutiny because of general unfamiliarity with non-probability methods. (See Section 7 for more discussion of the quality of estimates for probability and non-probability samples.)

Bias. Bias is an important quality measure for all surveys, regardless of how the sample was selected. Bias is the difference between the average of the estimates and the population values. While this definition applies to both probability and non-probability methods, the different methods used to make inferences require different ways of defining how the averages are computed.

In probability sampling, the design-based properties of the estimates are based on the theoretical average of the estimates computed over all possible random samples that could have been selected from the sampling frame under the chosen sample design. In computing this average, each of the estimates is weighted by the probabilities of selecting the sample. With a simple random sample, the probabilities of selection are identical so the average is just the unweighted, arithmetic mean. The computations are more complex with stratified and clustered designs, but these are all well described in sampling texts (Kish 1965, section 1.3; Valliant, Royal, and Dorfman 2000, section 1.3). Thus, the bias is the difference between the estimate averaged over all possible samples and the target population value.

The lack of a sampling frame and selection probabilities when using non-probability sampling render the design-based average over all possible samples infeasible. The Office of Management and

Budget (OMB 2006) uses the term *estimation error* to express the same idea for non-probability samples (see, e.g., Levy and Lemeshow 2008, section 2.4). The averaging schemes used to compute bias depend on the non-probability method used. For example, most model-based statistical approaches in standard statistics textbooks compute averages based on an assumed model for the characteristic being estimated. They may assume the characteristic, say income, follows a statistical distribution like the Normal distribution with a specified mean and variance, and the average is computed over this distribution. Assumptions that rely on formal statistical distributions are often unnecessary, and all that is needed is to assume the observations come from a distribution with a finite mean and variance. Once the averaging method is determined, bias is still defined as the difference between the average estimate and the target population value.

In addition to the design- and model-based methods, another method of making statistical inferences is called the *Bayesian method*. A recent example that discussed making inferences from a non-probability, online electoral poll used Bayesian methods. AAPOR (2012) released a statement on this. The Bayesian approach differs in important ways from the design-based and model-based methods described above. One key difference is that both the design-based and model-based methods start with the assumption that the value being estimated is a constant; for example, the total number of adults who are employed is a fixed number. In the Bayesian methodology, this total is assumed to be a random variable and inference involves using sample data to better describe the distribution of the total. To accomplish this, Bayesians begin with an initial or prior distribution of the total number of employed adults. This prior distribution may often be uninformative in the sense that it assumes we know little about the distribution. Sample data are then collected and the initial distribution is modified using Bayes' theory to produce a posterior distribution that incorporates what was learned from the sample. For example, the posterior distribution for the total number of employed might be approximated by a Normal distribution. The mean and variance of the posterior distribution provide information about the total number employed. A summary of the

posterior distribution of the region where the posterior distribution is most concentrated is called a *credibility interval.*

Although the Bayesian approach is very different from design-based and model-based methods in its representation of the outcome in terms of a distribution rather than an estimate of a fixed number, there are many similarities in practice. We do not discuss Bayesian methods further in this report, primarily because the work in applications to non-probability samples for surveys is very limited. We suspect that Bayesian methods could be applied to non-probability samples as successfully as other methods given further experimentation and development.

The goal in sampling is to produce estimates with small levels of bias where "small" can be interpreted differently depending on the purpose. In general, estimators with biases that get smaller as the sample size increases (consistency) are greatly preferred. In actual use, a small bias might be defined more practically. For example, Olson (2006) defined a small bias as a value that is not statistically different from zero in a nonresponse bias analysis. Others declare only substantively meaningful biases as being important.

Bias as defined above is not only due to observing only part of the sample, but it also includes nonsampling errors, such as those due to measurement error. The estimates that are being averaged in the computations are the actual estimates, not some theoretical values associated with the population.

Non-probability samples also have an important source of error that does not occur in probability samples (or at least should not occur) called *selection bias*. For example, in some quota samples interviewers might be asked to choose respondents with a specific age and sex composition, but otherwise the selection is left to them. The choices of the interviewers generally lead to selection bias (see, e.g., Lee 2006). Other non-probability sampling methods such as accepting volunteers will also be subject to selection biases (Bethlehem 2010). An important strength of probability samples is that they avoid this bias in the sampling stage, although it may arise in the response stage in both probability and non-probability samples.

Bias is a critical property of sample estimates because it can greatly affect the validity of those estimates. For example, Cochran (1977) showed how biases can lead to confidence intervals that include the true population value at much lower levels then the stated confidence level. Furthermore, biases due to nonsampling errors often do not decrease as the sample size gets larger, and confidence intervals for biased estimators have even lower coverage rates with large samples than with small samples.

Variance. The variance of an estimate is the average squared difference between the estimate and its average. The variance of an estimate is low if there is little variability among the estimates, even if none of the estimates is close to the population value being estimated. The variance is estimated as the weighted average of the squared difference between the estimate and its average overall all possible samples for probability samples in the design-based approach. For non-probability samples, the same averaging mechanism used in estimating the bias is used for the variance.

Mean Square Error. The mean square error of the estimate is a measure of accuracy rather than variability. As such, it is generally the preferred metric for comparisons across sampling methods. It is estimated identically for probability and non-probability samples; it is the squared bias plus the variance, but the bias and variances are estimated through different averaging mechanisms for probability and non-probability designs.

6.2 Estimation Procedures

The broad set of estimation procedures that have been used in non-probability samples can be classified into (pseudo) design-based and model-based. Both categories are discussed below.

Pseudo Design-Based Estimation. The statistical foundation of the design-based approach is that the one and only random component in the survey estimation process is associated with the probability of being selected into the sample. This random selection process is the basis for the evaluation or averaging over samples to compute the bias, variance, and mean square error of the

estimates. The inference to the target population is based on the random sampling procedure, not any randomness in the variables being estimated.

A *pseudo design-based weight* is sometimes used in non-probability samples. The term pseudo is added because unlike traditional design-based estimation, the selection probabilities are unknown and undefined (there is no explicit link from the sample and the frame, even if a frame exists). Instead an estimated probability of being in the non-probability sample is used instead of the known probability. The idea relies on the heuristic that each sample observation represents other non-sampled (or not responding) units. Once the pseudo weight is formed, other estimators, such as ratio estimators, are computed by substituting the estimated non-probability sample weight in place of the known probability sample weight in the traditional design-based sample.

The method of estimating the probability of selection differs from application to application, but a common theme is to estimate the probability by computing the ratio of the sample size to the estimated population total within some categories (e.g., poststratification). Clearly, the creation or estimation of a pseudo weight requires strong assumptions for some non-probability sampling recruitment methods.

Model-Based Estimation. Model-based estimation relies on a statistical model that describes the variable being estimated in the survey such as a Normal distribution. With this type of estimation procedure the characteristic of interest (the *y* variable) is assumed to be a random variable with a distribution, so the randomness does not come from the process of generating the non-probability sample. When the respondents are observed, the observations are used to fit the model and the analysis is conducted assuming the sampling can be ignored. In other words, the outcome estimated from the model is not statistically related to the method of sampling. This is similar to standard text-book statistics courses that assume the data come from a distribution like the Normal distribution and the data observed are used to estimate the mean and variance of the distribution so various estimates can be computed. A typical use of this approach requires that the observations are independent draws from the distribution of interest, an assumption that is violated if, for example,

we are interested in the characteristics of all Americans but only the residents of a single state are measured.

The statistical models often do not require imposing a specific statistical distribution. Perhaps more importantly, the models typically condition on covariates or auxiliary variables (e.g., one assumption is that the observations come from the same distribution when the covariates are the same). Several researchers have discussed this model-based analysis of survey data and the conditions under which the sampling method can be ignored and still make valid inferences (e.g., Little and Rubin 2002; Sugden and Smith 1984; Pfeffermann and Rao 2009).

The *Heckman two-step model* (Heckman 1976) is an example of a model-based method from the econometrics literature. This approach explicitly models the selection mechanism and the outcome variable using regression models at each step. The first step relies on the existence of an instrumental variable to help adjust for any selection bias in the variable of interest. The instrumental variable must be highly associated with the analysis variable but should not be subject to the selection bias that is affecting the analysis variable (see, e.g., Fuller 1987). Unlike more traditional measurement error models, the instrumental variable in the Heckman approach is a propensity of being in the sample estimated from an initial regression model. These estimated propensities are then used in the second step model to correct for the selection bias. Thus, the regression models for these two steps are correlated. Several strong assumptions (e.g., normality, correct model specification) are required for the method to produce efficient estimates; minor violations of the assumptions can produce unstable or biased estimates.

6.3 Weighting Adjustments

In probability samples, weighting is used to implement an estimation formula given a set of responses from a survey. Weights for probability samples begin with the base weights (sometimes called the design weight or inverse probability of selection weight). Weight adjustments are then applied to improve efficiency or to address potential biases, where the biases may be due to

nonresponse and coverage errors. Kalton and Flores-Cervantes (2003) described adjustments intended to reduce nonsampling errors in probability samples.

Non-probability samples may or may not use weights, but when they do use weights it is only a practical device and does not (or should not) imply that the methods used are design-based procedures. Clearly, base weights that are the inverse of the sampling probabilities cannot be computed for non-probability samples because these probabilities do not exist. Despite this problem, several approaches to produce estimates have been proposed for non-probability samples that involve weighting in one form or another. We summarize these below.

No weights. Econometrics and psychometrics, to name two fields not often thought of in the context of survey research, use questionnaires for collecting data. In these fields and in mathematical statistics, samples are typically assumed to be a random representation (i.e., simple random sample or SRS) of the target population (see, e.g., Casella and Berger 2002). Under the SRS assumption, no weights are needed to produce statistics such as means. Essentially, this is a form of pseudo weights where all the pseudo weights are set to a constant.

More complicated estimators use mathematical models that contain the key covariates as a way to reduce the reliance on the strong assumption of SRS. The viability of the SRS assumption for non-probability sampling is questionable in general because of uncertainty in how potential respondents are identified and sampled.

Propensity Score Adjustments. Propensity score adjustments (PSA) is one of several weighting methods that attempts to remove bias in both probability and non-probability samples. As discussed more extensively in Section 4, PSA were first introduced for observational studies (Rosenbaum and Rubin 1983) to try to balance known sample characteristics for comparison groups (e.g., treatment and control) after assignment to the groups has already been accomplished.

Propensity score adjustments (sometimes referred to as logistic regression weighting) are used extensively in probability surveys to limit nonresponse biases under the assumption that response is a random phenomenon (Little 1986; Little 1988; Fricker and Tourangeau 2010). In this application,

response models are developed using logistic regression that use predictor variables known for both respondents and nonrespondents. The resulting response propensity estimates (i.e., conditional probability of response) are used to adjust the base weights of the respondents by a factor that is assumed to be their probability of responding to the survey. Technical information can be found in the citations given above and section 11.2.3 of Bethlehem, Cobben, and Schouten (2011).

PSA methods also have been used in non-probability samples, especially with opt-in panels. They sometimes are proposed to adjust for the combined effects of coverage errors, nonresponse, and non-probability sampling. An extensive list of citations may be found in Lee (2006). To estimate the conditional probability of response across all these sources, a reference survey from a probability sample may be required. Using data from both samples, a logistic model is used to estimate the probability of participating in the non-probability study. In an ideal setting, the reference survey relies on a random sample selected from a frame that fully covers the target population with no nonresponse or other bias problems (Bethlehem and Biffignandi 2012, Chapter 11). However, many such PSA adjustments have included relatively small random digit dial (RDD) landline surveys with relatively low participation rates and coverage problems due to people switching to cellular telephone use only (Smith 2011). The potential biases from the reference survey could account for problems noted in the application of PSA to non-probability samples (see, e.g., Schonlau et al 2003).

When a good reference survey is available for the PSA the focus then turns to the model covariates or predictor variables. Model covariates may include: demographic characteristics that are the source for many poststratification adjustments; common attitudinal questionnaire items known as *webographic* characteristics[1] (Duffy, et al 2005); and observational and process data obtained during data collection known as *paradata* (Groves and Lyberg 2010). Research to date on the effectiveness

[1] Webographics are attitudinal variables thought to account for the differences between people who do surveys online and those who do not. They generally measure lifestyles issues such as the types of activities people engage and their frequency, media use, attitudes toward privacy, and openness to innovation.

of webographic questions (see, e.g., Lee 2006; Schonlau, van Soest, and Kapteyn 2007) and paradata (see, e.g., Kreuter, et al 2011) for propensity modeling has not been very successful.

Valliant and Dever (2011) examined the assumptions underlying the use of a PSA for non-probability samples. Their findings suggest:

- The probability of being included in the non-probability sample and of responding to the survey can sometimes be effectively modeled using variables collected in both surveys.
- The PSA should be generated from a model that incorporates the reference survey weights. Initial weights for the non-probability cases are typically set to one so that they only represent the respondent sample. However, some have suggested that a more appropriate approach would be to poststratify the non-probability weights to population counts prior to running the model (e.g., Loosveldt and Sonck 2008).
- The bias of any survey item that is associated with the actual probability of response but is available only for the non-probability sample will not be corrected using this technique. Unless the commonality of the questionnaire items for the non-probability and reference surveys is large, selection bias may affect a number of items collected for the non-probability sample. This finding may shed light on the mixed results on levels of bias found in the research to date (see, e.g., Malhotra and Krosnick 2007; Loosveldt and Sonck 2008).

Calibration Adjustments. Weight calibration has been studied extensively for probability samples and has been shown to reduce both bias and variance in survey estimates (Deville and Särndal 1992; Kott 2006). Two well-known and widely used examples of calibration are poststratification and raking. Poststratification is a very popular method in both probability and non-probability samples.

Another calibration method that is popular in Europe and gaining ground in the U.S. survey research literature is known *as generalized regression (GREG) weighting.* Through a linear regression model, calibration weights are constructed using a vector of auxiliary variables that are known for each element in the sample and have known population totals. The calibration weights are calculated in such a way that the survey estimates using these weights equal the known population totals for each of the auxiliary variables. The adjustment can be written as a weight that is a function of linear regression coefficients when linear calibration is used. The original approach suggested by Deville

and Särndal (1992) was intended to adjust the base weights in probability samples to improve precision. Later, it was expanded to include nonresponse and coverage adjustments.

Calibration methods traditionally are applied to quota samples, often weighting the observed quota samples so that they equal population totals for demographic variables such as race, age, and sex. The particular form of calibration used most often has been poststratification. In many non-probability samples poststratification is the only form of weighting, whereas in probability samples calibration typically is an additional adjustment of the weight formed using base weights and perhaps nonresponse adjusted weights. This type of weighting methodology may be the only useful tool for non-probability samples that do not have sufficient information to construct a PSA (samples without a corresponding reference survey) and those that do not control the sampling, such as used in matched sampling and network sampling.

Dever, Rafferty and Valliant (2008) found some, albeit inconsistent, benefits from GREG adjustment in reducing non-coverage bias. Citing problems with PSAs, Yeager, et al (2011) found that poststratification improved the accuracy of the non-probability sample estimates, though again their results were unstable. Research conducted by Lee (2006) and Lee and Valliant (2009), showed that either PSAs or calibration alone are generally not sufficient to reduces biases in the estimates from the non-probability surveys to relatively low levels.

Tourangeau and his co-authors (2013) summarized the results across eight studies that attempted to reduce biases in non-probability opt-in panels by using weighting methods when the biases were due to coverage and selection effects. They found:

- The adjustments removed only part of the bias, at most around three-fifths.
- The adjustments sometimes increased the biases relative to unadjusted estimates, sometimes by factors larger than two.
- The relative biases that were left after adjustment can be substantial, often shifting the estimates by 20 percent or more.
- There were large differences across variables, with the adjustments sometimes removing the biases and other times making them much worse.

Overall, the adjustments reduce to some extent, but do not by any means eliminate, coverage, nonresponse, and selection biases inherent in opt-in panels..

Other Weight Adjustments. Elliott (2009) devised a scaled "pseudo-weights" approach for combining a non-probability and probability sample. His method differs from using the probability sample as the reference sample in the PSA discussion above. His goal was to combine the two surveys and analyze them as one study. Elliott and Haviland (2007) further refined this methodology. They created composite estimates combining both the probability sample and the non-probability sample, where the non-probability sample estimates were computed using pseudo-weights. Their estimator is similar to methods used for Bayesian estimation where estimates are combined in a way that gives more "weight" to the more precise estimate. The specific estimator that they use for this purpose is one proposed and used in small area estimation applications (see, e.g., Rao 2003). They showed that their proposed estimator had a smaller mean squared error than a similar estimator generated from only the reference sample.

6.4 Variance Estimation

Variance estimation in probability samples uses a design-based approach based on the probabilities of selection and averaging over all possible random samples with the same design. With non-probability surveys, less extensive research on variance estimation has been published, as most of the interest has been directed at bias. However, more research in this area is essential as variance estimates are needed for inferences and for evaluation of the estimates. In the absence of such research, statements such as those of the National Agricultural Statistical Service (USDA 2006) that discuss the inability to make valid, design-based estimates of variability from non-probability samples are sometimes interpreted to mean that non-probability samples cannot support any variance estimation approach.

An approach often taken is to assume a SRS design and estimate a standard deviation under this assumption. This approach sometimes is also used for probability samples, but the literature is very

clear that ignoring the sample design produces biased estimates of precision. The same is undoubtedly true for most non-probability samples.

Because design-based variance estimates cannot be made from non-probability samples, techniques have been explored to calculate *error variance* for non-probability surveys. Thompson (1990, 2002) described pseudo design-based methods for specialized adaptive designs. Isaksson and Lee (2005) suggested possible techniques involving "propensity score poststratification." Finally, de Munnik, Dupuis, and Illing (2009) described a resampling technique.

Pseudo Design-based Methods. Thompson (1990, 2002) described modified "Hansen-Hurwitz" and "Horvitz-Thompson" estimators which account for the inclusion probabilities of units captured through their connection with the original random sample. The authors suggested using a (stratified) SRS formula with the modified values for the estimated variance. Note that this SRS approach has been adopted by others after incorporating the estimated weights as discussed above.

Propensity Score Stratification. The term *propensity score stratification* is a technique where the estimated selection probabilities are used in poststratification in addition to the analysis weight. A poststratified variance estimator was used with two of the three approaches examined by Isaksson and Lee (2005). The first and second approaches used poststratification methods with one including a model-based modification. The third variance estimator randomly divided the sample into groups of equal size and then a standard jackknife variance estimator was used on the replicate estimates. This last approach overestimated the true variance.

Resampling. Bootstrap methods have been used in many contexts to estimate variances (Efron & Tibshirani 1993). The general approach is to draw random subsamples from the sample a large number of times, estimate the statistic of interest from each of these subsamples, and then use Monte Carlo approximation methods to estimate the variance of the estimates. For probability samples, Rao and Wu (1988) proposed a bootstrap approach that can be used for sample surveys.

The bootstrap approach was used but not evaluated by de Munnik, Dupuis, and Illing (2009) for non-probability samples.

6.5 Summary

The main concern with non-probability samples is that population estimates may be highly dependent on model assumptions. The assumptions may range from simple to complex, and the model may be implicit rather than explicit. When the model assumptions are reasonably good approximations then non-probability estimates behave well in terms of having the expected levels of (low) bias and variance (Valliant and Dever 2011).

The difficulty is in knowing when the models are good approximations. The missing link (the selection probabilities) between the sample and the target population complicates matters. This link for probability surveys is used to describe those population units not available for sampling (undercoverage) and characteristics related to nonresponse bias among those who could have been captured by the sampling procedures.

The task of quantifying the quality of the non-probability survey estimates is daunting. Techniques already established for probability sample surveys are being used for this purpose, and some new techniques have been devised specifically for non-probability studies. As the field continues to expand these techniques, the situations in which non-probability surveys may be most appropriate may become clearer.

In the meantime, there are a wide variety of approaches being investigated or already in use in fields often not familiar to opinion, social and market researchers. With the growing awareness of the advantages of non-probability samples in terms of cost and speed we expect, or at least hope, that will change. In this section we have discussed many of these approaches, although only at a very high level.

7. MEASURES OF QUALITY

Measuring the quality of data from non-probability samples is a new and challenging task. For decades survey researchers have relied on measures for assessing data quality based in the probability sampling paradigm. Many survey organizations and government agencies have established standards and guidelines to produce these measures, and they are often documented in quality profiles[2] that summarize what is known about the sources and magnitude of error in data for the surveys they conduct. Because they are grounded in probability theory and have been used for decades, they are widely accepted as useful measures of quality.

Unfortunately, non-probability samples violate three key assumptions on which many of these measures are based. Those assumptions are: (1) a frame exists for all units of the population; (2) every unit has a positive probability of selection; and (3) the probability of selection can be computed for each unit. The standard quality metrics are designed to measure the degree to which a specific sample violates these assumptions due to such real-life constraints as incomplete coverage and unit nonresponse.

The quality standard guidelines issued by Statistics Canada, the U.S. Office of Management and Budget, and the U.S Census Bureau contain little commentary on methods for assessing data quality in non-probability samples. For example, Statistics Canada (2009) acknowledges that non-probability samples can be an easy, fast and inexpensive way to conduct preliminary design studies, focus groups and follow-up surveys. But they also go on to say, "The ability to make reliable inferences about the

[2] Quality profiles have largely been done for government surveys. As an example, one is available at http://www.census.gov/sipp/workpapr/wp30.pdf.

entire population and to quantify the error in the estimates makes probability sampling the best choice for most statistical programs."

The U.S. Office of Management and Budget (2006) has issued standards and guidelines for Federal statistical surveys. The document describes 20 standards that cover in great detail all stages of a survey including design, pretesting, data collection, data processing and editing, data analysis, and data dissemination. However, they offer almost no guidance on non-probability samples. Agencies are instructed to select samples using accepted statistical methods, for example, "probabilistic methods that can provide estimates of sampling error" (OMB 2006, pg. i). The use of non-probability samples must be "justified statistically and be able to measure estimation error". To wit:

> When a nonprobabilistic sampling method is employed, include the following in the survey design documentation: a discussion of what options were considered and why the final design was selected, an estimate of the potential bias in the estimates, and the methodology to be used to measure estimation error. In addition, detail the selection process and demonstrate that units not in the sample are impartially excluded on objective grounds in the survey design documentation. (OMB 2006, p. 7).

The statistical quality standards issued by the U.S. Census Bureau are equally unhelpful. Requirement A3.3: reads "Sampling frames that meet the data collection objectives must be developed using *statistically sound* methods" (italics added). The requirement goes on to say, "Statistically sound sample designs require a probability sample" (U.S. Census Bureau 2011, pg.25).

As we have noted elsewhere in this report, virtually all non-probability methods must overcome three obstacles: the exclusion of large numbers of people from the sampling process; reliance on volunteers or referrals; and, in many instances, high nonresponse. Some methods, such as convenience sampling, largely ignore these problems while others, such as sample matching and post-survey adjustments of various kinds, go to considerable length to identify the right set of auxiliary variables that can be used to adjust the sample so that it is representative (or at least approximately

representative) of the target population. However, as of this writing there are no widely-accepted measures or practices for validating the assumptions and the efficacy of those adjustments.

The absence of these measures and practices may be the primary reason why survey researchers, especially those charged with producing accurate national benchmarks, have been reluctant to accept non-probability methods. And indeed, AAPOR recently issued a statement noting that the difficulty of validating the assumptions that underlie claims of representativeness for opt-in panels causes it to continue to recommend probability sampling when precise estimates are needed (Baker et al. 2011, p. 758).

Without an accepted framework for assessing data quality in non-probability samples survey researchers might still find it useful to conceptualize the problem within the familiar framework of Total Survey Error (Groves, 1989). This framework groups errors into two broad categories: errors of non-observation and errors of observation. The former describe errors that affect the representativeness of the sample and the latter errors that affect the measures of interest because of how the survey was designed and executed. Given that the most frequent criticism of non-probability samples is that they are not representative, we consider errors of non-observation first.

7.1 Errors of Non-Observation

Errors of non-observation derive from gaps between the target population, the sampling frame, and the sample. They generally are grouped into three categories: coverage error, sampling error, and nonresponse error.

Coverage Error. Coverage error measures how well a sample frame covers the target population. Under ideal circumstances every member of the target population is listed in the sampling frame and therefore has a chance to be selected for the sample. Groves et al. (2009, pg. 84) have argued that "without a well-defined sampling frame, the coverage error of resulting estimates is completely unknowable." Since in non-probability samples units generally are not sampled from a well-defined sampling frame, the task of assessing coverage error seems impossible.

However, this assessment is deeply rooted in the probability sampling paradigm and may not be directly applicable to non-probability samples. In probability samples, coverage ratios are the most commonly used metric of coverage error. A coverage ratio is the ratio of the estimated total from the sample computed by using only the inverse of the probability of selection weights (perhaps adjusted for nonresponse) to the known population total. In household surveys, these are typically demographic ratios where the denominators are data from recent censuses.

Non-probability samples can compute similar ratios, but they do not begin with inverse sampling weights so the computations must be different. It is possible to compare relative ratios from a non-probability sample to population totals from a census. For example, suppose the target population of a study is intravenous drug users within a county and a respondent-driven sample is used to locate and recruit respondents for this hard-to-reach population. If the health clinic records that serve that county's population are assumed to provide reasonably good population controls, then demographic ratios such as sex ratios by age group and race/ethnicity category could be computed. If the demographic ratios of the survey are vastly different from those reflected in the records, there may be evidence of coverage error. If the ratios are similar, however, it does not imply that the coverage is complete. It only implies that the coverage is similar across the different age and race/ethnicity groups.

The principal problem is that non-probability samples that attempt to estimate totals must rely on external population totals to produce the equivalent of sampling weights (since they are not sampled from a frame). This makes the computation of absolute rather than relative coverage ratios difficult. Researchers have used different ways to compute absolute coverage ratios, but with at best mixed success. For example, one approach for online surveys using sample from opt-in panels is to estimate the proportion of the population that has access to the Internet as a coverage measure. While this is somewhat informative, it makes an assumption that all those who do have access to the Internet are exposed to the survey request and therefore eligible to participate. This simply is not true

in almost all cases we can think of. More direct measures of coverage are needed for non-probability samples.

Of course, coverage ratios, much like response rates, are imperfect measures of biases due to non-coverage even in probability samples. As an extreme example, suppose the sampling frame of addresses in the postal service delivery files contains the address of 98 percent of the entire U.S. population and it is used to draw a probability sample. The coverage ratio is very high, but this sample would have dreadful coverage for estimating characteristics of the homeless.

A recent article by Blair and Conrad (2011) discusses the notion of "problem" coverage in the context of cognitive interview pretests, and focuses on how to select a non-probability sample size that is adequate to detect questionnaire problems. For probability samples, the notion of sample size is directly related to the level of sampling error one can expect – the relationship for non-probability samples is much less clear. Blair and Conrad provide a first attempt to quantify and give practical guidelines for sample size determination in this setting. They demonstrate that different sizes of quota samples (ranging from as few as 5 to as many as 90) are needed to observe different outcomes (e.g., mean number of problems discovered, likelihood of high impact problem discovery). One of their conclusions is that "small samples sizes may miss a substantial percentage of problems, even if concern is limited to those problems with a serious impact on measurement error" (Blair and Conrad 2011, pg. 654). More empirical studies such as these are needed to guide quality considerations for quota and other non-probability samples.

Sampling Error. The next component of the total survey error model, sampling error, measures variations in the survey estimates over all possible samples under the same design utilizing the same sample frame. When the sample frame is assumed to provide full or nearly full coverage of the entire population, then it is reasonable to assume that such measures are at least an approximate estimate of the precision of the survey's estimates. However, full coverage is seldom a viable assumption with non-probability samples. The case is especially clear with non-probability panels

and so AAPOR has long maintained "reporting margin of sampling error with opt-in or self-identified samples is misleading" (Baker et al. 2011, p. 773).

We agree that margin of sampling error in surveys has an accepted meaning and that this measure is not appropriate for non-probability samples. However, the broader statistics literature does use terms that are directly comparable to sampling error as a measure of variation of the estimates that is not tied to the idea of all possible samples. For example, most elementary statistics texts assume a model (suppose a random sample is selected from a Normal distribution) and estimate the standard error of a statistic with respect to that model as a measure of the uncertainty due to sampling.

We believe that users of non-probability samples should be encouraged to report measures of the precision of their estimates, but suggest that, to avoid confusion, the set of terms be distinct from those currently used in probability sample surveys. The precision of estimates from non-probability samples is not the average deviation over all possible samples, but rather is a model-based measure of deviation from the population value. Ipsos, for example has proposed the *credibility interval* (Ipsos, 2012) for their estimates from an opt-in panel survey. As noted in Section 6, the credibility interval is measure of uncertainty that is used with Bayesian methods, and Ipsos described their procedure as Bayesian. Other model-based approaches also produce estimates of precision such as standard errors that could be used and do not refer to the average over all possible samples (the accepted terminology for design-based inferences used in probability samples).

Although the research base does not exist to endorse this particular measure or to urge its adoption across the industry, we believe the industry needs constructive attempts to develop measures that fills the gap created by the unsuitability of the standard margin of error calculation with non-probability samples. Treating estimates as though they had no error at all is not a reasonable option. At this point, it falls to individual researchers to judge the usefulness of this particular measure. Such judgments are only possible when organizations using them fully disclose

the full range of information specified in the AAPOR Code of Professional Ethics and Practice along with a detailed description of how the underlying model was specified, its assumptions validated, and the measure calculated.

There are, of course, many methods for selecting non-probability samples and so no single approach to estimate a measure of error is likely to be appropriate. For example, Sudman (1966) proposed a technique for quota sampling and showed that if the sample is selected using the procedures he outlined and certain assumptions are satisfied estimates of sampling error like those used with probability samples also can be used with quota samples selected under his method. Other methods of estimating sampling error have been developed for respondent-driven sampling (see Section 4). Once again, because the methods used to draw non-probability samples are so diverse, it is incumbent upon those who use these methods to clearly document their methods and the assumptions underlying their computations of variability or uncertainty.

It is important that non-probability sampling methods be evaluated more closely with respect to any claims about their measures of precision. One well-established approach to accomplish this is to use replicate samples, an idea that traces back to the early days of probability sampling (Mahalanobis, 1946; Deming,). This approach is also used in observational studies and experiments.

In its purest form, multiple samples are selected and the variation in the estimates across these samples is a measure of the standard error of the estimate. *Pseudo-replication* methods take this a bit further and take one sample and divide it into replicate samples. The theory for pseudo-replication in probability samples is well established (Wolter, 2007), and developments for non-probability methods are feasible. Replication can be economical, relatively robust, and understood by many practitioners. It does not require larger sample sizes and can be executed relatively quickly. However, like traditional margin of error computations in probability samples, this approach only estimates variance; biases in the estimates cannot be measured.

Nonresponse error. The response rate in a probability sample is the ratio of eligible units measured to the total number of eligible units in sample (AAPOR 2009). The response rate is probably the most recognized quality measure for probability samples, even though in recent years its limitations as a predictor for nonresponse bias of estimates have been made clear (Groves 2006).

In non-probability samples, the denominator for the ratio may not be known, therefore it is not always possible to produce response rates as traditionally defined by AAPOR and other professional standards bodies. Consequently, as with margin of error calculations, researchers reporting on non-probability samples should avoid the term "response rate" and instead use another term. 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). This term has been adopted by AAPOR and is included the 2011 revision of its Standard Definitions. Eysenbach (2004) suggests a number of other related measures including *view rate* and *completion rate*.

Opt-in panels often collect detailed profiles as part of the recruitment stage that researchers might use to assess differences between the responders and non-responders to specific studies. They might make use of this information to assess potential nonresponse bias and data quality, but as of this writing we are not aware of any instances in which this is being done.

Other response metrics specific to opt-in panels are covered in detail by Callegaro and DiSogra (2008). Some, but not all, apply to non-probability panels. These include:

- The a*bsorption rate* measures the degree to which email invitations are not delivered to members because of wrong addresses, full mailboxes, or a network delivery error. This measure is one indicator of how well a panel provider updates and communicates with its members.
- The *break off rate* is the number of surveys that do not meet a preset threshold of answered questions. High break-off rates can be an indicator data quality by signifying questionnaire design problems (e.g. poorly designed formatting or navigation or indication the survey is too long).
- The *screening completion rate/study-specific eligibility rate* is the number of people who complete the screener and qualify plus those who complete the screener but do not qualify divided by the

total number of survey invitations. If screening and eligibility rates are very different from well-established external benchmarks, this may indicate fraudulent/incorrect reporting by panel members who are purposefully self-selecting into studies.

- The *attrition rate* measures the percentage of members that drop out over a set period of time. It is measured by counting the number of panel members that remain active month after month (Clinton 2001). High attrition rates may signal poorly designed questionnaires (e.g. surveys that are too long) that result in panel fatigue and high dropout rates.

The relatively few measures that have been developed to evaluate online samples mostly assume an opt-in panel of the sort that has dominated online market research over about the last decade. But the panel model is rapidly falling into obsolescence as the demand for online respondents grows, clients look for greater demographic diversity in their online samples and interest in studying low incidence populations increases. Providers of online samples are increasingly relying on a range of sources that expands beyond their proprietary panels to the panels of competitors, social networks, and the use of general survey invitations placed on a variety of websites across the Internet, much like river sampling. These respondents may no longer receive invitations to do a specific survey on a specific topic. Instead they receive a general solicitation to do a survey that directs them to a website where they are screened and then routed to a waiting survey for which they already have qualified. The software that controls this process is called a *router*. Its goal is to ensure that anyone willing to do a survey online gets one. As of this writing there is a good deal of variation in how routers are designed, how they operate, and what impacts, if any, they have on the data. Unfortunately, they also make it impossible to calculate a participation rate as we have discussed it above.

Nonresponse bias studies have become part of most government-sponsored probability sample surveys with OMB playing a key role in the development of nonresponse bias analysis approaches. The use of subsamples of nonresponders may be applicable to specific types of non-probability sample surveys. Techniques that make use of individual/geographic internal characteristics available in the sampling frame are widely used in probability sampling and may be applicable to non-probability samples that collect information on the pool of potential respondents. In some cases external benchmarks are available for one or more key substantive variables included in the survey.

When little internal or external information is available it may still be possible to assess nonresponse bias using face validity methods. For example, the medical literature may indicate that roughly half of the persons with a specific condition are females. If one uses a non-probability sampling technique to produce a sample of persons with the condition, using say nonrandom seeds, and 80% of the respondents are females, then it is clear that in the process of generating the sample males were much less likely to be nominated and/or participate in the survey. In this situation it is not clear that weighting the sample by gender and other socio-demographic can account for the 30 percent of males that are missing from the sample.

Reinforcing the complexity of non-response in non-probability samples, Gile, Johnston and Salganik (2012) introduced a three-level measure of non-response for respondent-driven sampling. Based on respondents' reports of the numbers of coupons refused by their contacts, and the numbers of coupons they distributed, the authors computed a coupon-refusal rate, and a coupon non-return rate, which combine to form a total nonresponse rate.

7.2 Measurement Error

The other broad category of errors accounted for in the TSE framework describes errors in observation, commonly referred to as measurement errors. These errors are generally seen to arise from four sources: the questionnaire, the interviewer (if there is one), the respondent, and the mode of data collection (e.g., face to face, mail, telephone or web). Here we might argue that data coming from non-probability samples likely have the same error properties as those coming from probability samples. Both are prone to observational gaps: between the constructs intended to be measured and the measurements devised to quantify them; between the application of the measurements and the responses provided; and between the response provided and the data ultimately recorded and edited. Whether the survey data come from a probability or non-probability sample should not matter --we should be able to use the same type of measurement error quality indicators.

The Questionnaire. One of the most common quality indicators associated with the survey questionnaire is *construct validity*, that is, the degree to which survey items that measure the properties of key constructs correlate in expected ways or the degree to which the survey has measured the true value of a construct. For example, if we wanted to assess data privacy or confidentiality concerns we would expect that respondents answering that they are "very concerned" about data privacy might also express concern about computer identity theft, hackers and belief that government agencies routinely share data. If we found these items to be correlated in the expected direction, then we would have some reassurance that the underlying construct is being measured. We might also compare key items across demographic or other defined groups that we would expect to respond differently on those items. For example, in a study of political attitudes on social issues such as same-sex marriage, gun control, and limits on abortion we would expect to see consistent differences between Republicans and Democrats.

A more direct measure of construct validity uses multiple items within the questionnaire specifically designed to replicate key items. The National Household Survey on Drug Use uses this technique to validate respondent reports of marijuana use (Biemer and Lyberg 2003). The first questionnaire asks, "How long has it been since you last used marijuana or hashish?" Later in the same survey, respondents are asked, "On how many days in the past 12 months did you use marijuana or hashish?" If a respondent answers "a day or more" to the second item then we would expect the answer to the earlier item to be something less than one year. High levels of inconsistency between these items might signal problems in the survey's measurement of marijuana use.

The Interviewer. The interviewer does not play a role in many of the newer non-probability sample surveys since self-administered paper or web surveys are commonly used. The interviewer as a source of measurement error is however relevant when non-probability sample surveys are conducted by telephone or in person. Again, there is an extensive literature on the interviewer as a source of measurement error in probability sample surveys (Groves 1989) and much or all of this

literature is relevant to non-probability sample surveys that use interviewers. These techniques include embedded designs for measuring interviewer variance, examining the relationship between interviewer characteristics survey responses, taking advantage of the feasibility of randomizing interviewers to experimental survey conditions in centralized telephone interviewing facilities, measuring interviewer compliance with training materials, and examining paradata to identify potential cases on interviewer falsification.

The Respondent. All non-probability sample surveys rely on a set of respondents generated in some nonrandom fashion and those respondents are a potential source of survey measurement error. Kahn and Cannell (1957) and Groves (1989) identified five stages of respondent reporting:

1. Encoding of information
2. Comprehension
3. Retrieval
4. Judgment of appropriate answer, and
5. Communication

Virtually all of the techniques to assess respondent reporting errors developed for probability sample surveys are relevant. Reverse record check studies can be used to compare respondent reports with administrative/program/medical records. Reverse record checks are also potentially very useful at the design stage to help determine the length of recall period for the reporting of events from memory. Cognitive interviewing can also be very helpful in assessing measurement error related to one of more of the above steps. Cognitive interviewing can be useful for testing respondent comprehension of eligibility screening criteria, especially if the eligibility characteristics of the cognitive interviewing participants can be accurately determined external to the screening questions being tested.

The widespread use of the web as a data collection modality allows the use of technology to obtain detailed analyzable information on the respondent interaction with the questionnaire. In addition to recording information on break-off points, well designed web questionnaires can also

allow for the measurement of time spent on each question, forward and backwards movement within the questionnaire, frequency of selection of "don't know" response categories as a measure of item nonresponse, and length (i.e., number of words) of open ended responses.

Finally, there are a set of quality issues that have their roots in the opt-in panel paradigm. These are discussed in 7.4.

The Mode of Data Collection. Groves (1989) among others has described the response effects most affected by the mode of data collection. Today, surveys are conducted in-person, in-person with the respondent provided with a communication device to call into a centralized telephone interviewing facility, using mail administered surveys, by telephone using decentralized telephone interviewing consisting of one interviewer making calls, by telephone from centralized telephone interviewing facilities, and using the web. The picture is even more mixed when we consider various hybrids such as in-person interviewing where the respondent completes a self-administered interview using a computer-assisted device.

Over the past 10 years there has been considerable survey modality research conducted for a wide range of probability sample surveys. The growth of the various types of multi-modality surveys has allowed for mode experiments to be embedded in probability sample surveys. The primary objective of these experiments is to determine whether the mode of data collection leads to mode effects. The literature on mode effects for mail surveys and web surveys is particularly relevant to non-probability sample surveys because these seem to be the two primary modes of data collection. Some types of non-probability sample surveys are amenable to randomization of mode to mail or web or the use of mode choice experiments whereby the respondent is given the option to select among two or more modes. It is also important to examine break-off rates in multi-modality surveys and to determine where break-offs are occurring in the screening/main questionnaire.

7.3 External Validity

The idea of external validity is central to much of the work done thus far that attempts to assess the quality of survey data collected from opt-in panels. A literature has developed over the last decade describing attempts to assess the validity of samples from these sources as compared to probability samples of varying quality (mostly on the telephone), benchmark data such as censuses, electoral outcomes and other data collected by non-survey methods such as administrative or sales data. This literature was reviewed in the earlier AAPOR task force on opt-in panels (Baker et al. 2011) and will not be reprised here. Arguably the most frequently cited work is the Yeager et al. (2011) evaluation of five opt-in panels which compared a small set of variables to two probability samples, a high quality government survey and administrative records.

On the face of it this would seem to be an effective way to assess the validity a specific non-probability sample. In practice it is problematic for at least three reasons.

First, it's not always practical because the external measures needed for comparison may not exist. This can be done in carefully designed experiments for this purpose, but this is not the standard approach to doing surveys. If the target population and the survey topic are frequently studied or monitored it may be more practical to find sources for comparison. But as target populations become more narrowly defined and topics more esoteric, such sources may be difficult if not impossible to find.

Second, even very high quality probability surveys that purport to measure the same thing sometimes don't agree. For example, the number and percent of persons who are uninsured is estimated in at least four U.S. federal government surveys, all of which have high response rates. Nevertheless, there are differences in the estimates that far exceed sampling error and those appear to be due to measurement and other nonsampling errors (Davern et al. 2011). Another example is the Survey of Income and Program Participation (SIPP) estimates of the number of persons with General Education Development (GED) test high school equivalencies to be about 70 percent

higher than the estimate from the Current Population Survey (CPS) School Enrollment Supplement (Crissey and Bauman 2012). Both sources of these estimates are high quality surveys conducted by the U.S. Census Bureau.

Finally, administrative records and other kinds of data collected by non-survey means are not error-free. For example, Smith et al. (2010) described a situation in which the misclassification of ethnicity and race in administrative records occurred in nearly 25% of the records, the majority due to missing information. This is not an unusual situation in administrative records that are not developed to support statistical uses. Brackstone (1987) reviewed the advantages and disadvantages of administrative records for these purposes. Boruch (2012) also assessed the uses of these types of records and provided a different and thought-provoking discussion.

The bottom line would seem to be that in many cases it is very difficult to determine the reason the estimates from different surveys differ and to assign some proportion of the difference to a specific cause like the sampling mechanism. Further, gold standards for comparison do not exist in most cases. Nonetheless, we can make judgments about validity assuming we understand something about the quality of the comparison data that are available or when multiple sources for comparison exist, but precise estimates of error due to the sampling mechanism are very difficult to come by.

7.4 Quality Measures Unique to Opt-in Panels

Over about the last 10 years the market research industry has come to rely on a variety of non-probability methods used in conjunction with online access panels. Some of these methods have been described in previous sections and results from studies using opt-in panels have frequently been challenged on grounds of both bias and variance (Yeager et al. 2011; Walker and Petit 2009). However, these and other studies like them generally have not looked at the specific sampling methods used. Their focus has tended to be on the panels themselves rather than the techniques used to draw samples from them. From that perspective it seems entirely legitimate to focus on coverage and in turn point out the variety of ways in which individual panels are recruited and

maintained. Likewise, calculating estimates of coverage error and examining the techniques that may be used to adjust for it are essential parts of evaluating the quality of samples from opt-in panels.

However, it is now widely accepted across the market research industry that there are three additional problems that are unique and probably endemic to the panel model. They are:

- The tendency for people to sign up on the same panel using different identities in order to increase their chances of being selected for surveys.
- The possibility that when more than one panel is used to get a significant number of lower-incidence respondents that some people may be sampled and complete the same survey more than once. A more malevolent form involves the use of web robots (or simply bots) to automatically complete the same survey multiple times.
- High rates of satisficing as evidenced by very short completion times, straightlining in matrix style questions, frequent selection of non-substantive responses, skipping questions, and inconsistent or nonsensical responses.

The magnitude of these problems across the industry and the degree to which they affect survey estimates is a matter of some debate, and one suspects that there is a good deal of variation from panel to panel if not survey to survey. Nonetheless, they are now widely acknowledged across the industry as significant quality problems to be addressed by industry and professional associations, software developers, panel companies and individual researchers. For example, both ISO 20252 – Market, Opinion and Social Research and ISO 26362 – Market, Access Panels in Market, Opinion and Social Research specify requirements to be met and reported on. The market now has two software solutions, RelevantID and TrueSample with the functionality to identify potentially problematic respondents in online samples. Many individual researchers have implemented post survey editing procedures designed to isolate duplicate respondents and potential satisficers (Le Guin and Baker, 2007).

But these sample quality procedures are not without problems of their own. For example, the procedures commonly used to validate the identity of prospective panelists have been shown to reject people under 30, the less affluent, the less well-educated and non-whites at a much higher rate than other demographic groups (Courtright and Miller 2011). Likewise, there are no clear standards

for measuring survey engagement or separating out poor respondent performance from poor questionnaire design.

Nonetheless, researchers interested in assessing the quality of samples drawn from opt-in panels are wise to insist on reports of the specific steps taken to ensure that all respondents are real people with the characteristics they claim to have; that no respondent was allowed to complete more than once; and that unengaged respondents have been identified and the measures used to engage them clearly defined. The measures used and the extent to which they provide a workable assessment of sample quality are not yet established standards, but they can provide the researcher another quality assessment tool when working with opt-in panels.

7.5 Other Metrics on the Horizon

As we noted at the outset, the two main measures used in probability sampling to evaluate the representation component of TSE, coverage ratios and response rates, may not be very good indicators of bias. Researchers who routinely work with probability samples have been investigating alternative approaches to deal with this shortcoming, especially with respect to response rates. For example, *R-indicators* are one approach to replace or supplement response rates (Schouten, Cobben, and Bethlehem 2009). Särndal (2011) also addressed this problem and suggested balancing the sample and the set of respondents to reduce nonresponse bias. Essentially, the R-indicators and some of the indicators proposed by Särndal and his colleagues involve comparing response rates for subgroups identified from auxiliary data. If the response rates from all the subgroups are relatively consistent, then there is little evidence of nonresponse bias (nonresponse bias occurs when response rates co-vary with the characteristics being estimated). The indicators quantify this variation in response rates across the subgroups.

Other methods have also been proposed that are more directly applicable to non-probability samples. Frank and Min (2007) have suggested a testing approach to assess whether causal inferences can be generalized when the data are not a probability sample of the entire population. Fan (2011)

proposed a method for constructing tolerance "maps" to provide users different tolerances of non-representativeness in different studies. Using data from Louis Harris and Pew Internet and American Life Project surveys, he demonstrated how representative responses might be obtained from a non-probability Internet sample recruited exclusively from politically conservative websites. In concept, this new method is consistent with the fit for purpose ideas discussed in the next section since it allows researchers to "devise samples capable of giving responses that are less than truly representative but that are still within the researcher's tolerance".

7.6 Summary

The general lack of well-defined measures for assessing the quality of non-probability samples is due in part to the lack of a single framework within which all non-probability methods fall and, perhaps more importantly, the willingness of practitioners historically to accept these methods at face value. Survey researchers, on the other hand, are accustomed to using widely accepted quality measures that they believe substantiate and qualify the inferences they make from probability samples. It seems clear that if non-probability methods are to be embraced as valid for surveys then similar measures or methods are needed.

The TSE model may offer some help. Errors of non-observation and especially coverage and nonresponse error pose the biggest challenge. Sampling frames for non-probability samples are seldom defined in ways that cover the full target population and give every member of that population the opportunity to be sampled. Measures of relative coverage may be possible in some instances, but because they are generally based on variables available for comparison (such as demographics) there is no guarantee that they measure the characteristics that matter most.

In the end, it may be more important to validate the steps taken to overcome coverage and nonresponse error in a non-probability sample than it is to measure it, at least for routine practice.

Establishing external validity or representativeness through comparison of survey estimates to benchmark data such as that from censuses, high quality probability sample surveys or administrative

records is sometimes possible, but it's not clear that the data for comparisons always exist and are at least reasonably error-free. Replication might also be helpful to assess the variability of the estimates and it can be relatively easy to do with at least some non-probability methods.

We think it reasonable to approach the measurement side of the model with pretty much the same tools as we use for probability samples. Where errors of observation are concerned we probably can expect that non-probability samples have error properties that are similar to probability samples using similar modes.

As we have said at different points and in different ways in this report, the inherent risk in model-based methods is that the assumptions underlying any given model do not hold and the outcomes are sensitive to those assumptions. When this happens, the estimates are biased, possibly severely biased. If we are to make informed judgments about the quality of a non-probability sample the underlying assumptions of any adjustments made must be clearly stated, empirical validation of those assumptions described and the auxiliary variables used defined. We welcome the development of new approaches and measures such as the credibility interval, but note that migrating such measures from the proposal stage to the acceptance stage requires more validation than we have seen to date. Even then, researchers accustomed to empirical measures rooted in strong theory and decades of practice must become comfortable making judgments that may not be black and white.

Transparency of method has long been championed by AAPOR as the only route to careful evaluation of the quality of a survey. This is especially true when the survey relies on a non-probability sample. However, there is no apparent consensus as to what should be disclosed. The AAPOR Code has a set of items that are appropriate to probability-based surveys but may not work well for sources such as opt-in panels. The Public Works and Government Services Canada (2008) has proposed a disclosure standard for online surveys and two ISO standards (20252 and 26362) have done so as well.

The bottom line is that the multiplicity of approaches and methods along with their continued evolution make judgments about the quality of non-probability samples difficult. And in most cases individual researchers and data users are forced to make case-by-case decisions about the quality and utility of estimates from studies that use these kinds of samples. To make progress, procedures for certain types of non-probability samples like opt-in panels need to stabilize and estimates from these samples across a broad range of applications need to be evaluated. If such results are consistent with a priori theoretical assessments of the accuracy or stability of the estimates, then support for these methods will be greatly enhanced.

8. FIT FOR PURPOSE

A key reason for the predominance of probability sampling over about the last 60 years has been its accuracy, that is, its demonstrated ability when its key assumptions are met to generate population estimates that are within a calculable range of the true values for that population. During much of that time, discussions about survey data quality generally have been framed using statistical concepts such as bias and variance with varying attempts to define the types of errors in survey design and implementation that produce them. Groves (1989) summarized decades of statistical and social science literature on total survey error (TSE), described the principal categories of error and their impact on bias and variance, and tied these to the costs of conducting the survey.

As discussed in the previous section, the TSE model is an excellent lens through which to view the sources of errors in probability-based surveys, with some potential usefulness in non-probability surveys as well. Nonetheless, Groves and Lyberg (2010) argue that TSE should be but one approach within a much broader quality framework. First, costs are important and must be considered in a practical sense. Second, a broader framework is needed to assess survey results in light of how those data are to be used, a set of criteria expressed in terms such as *fit for purpose* or *fitness for use*.

8.1 The Changing Definition of Quality

In their book, Biemer and Lyberg (2003) linked changing definitions of survey data quality to the broader quality revolution of the 1980s and 1990s. In manufacturing settings the definition of quality once amounted to something like "free of defects" or "conformance to specifications." With advent of the Total Quality Management (TQM) movement this changed and one begins to see a stronger emphasis on the somewhat subjective needs of the customer and the use to which a product was to be put. As Deming (1982) writes "Quality should be aimed at the needs of the consumer." Juran (1992) is often credited with having invented the term, *fitness for use,* by which he meant that how a product would be used and the price a customer is willing to pay should be important factors in the design process and therefore an essential part of the concept of quality.

ISO, the International Organization for Standardization, has as its mission the development and dissemination of international standards for products and services. Working through a network of national institutes in 162 countries, ISO has developed over 19,000 global standards to "ensure desirable characteristics of products and services such as quality, environmental friendliness, safety, reliability, efficiency and interchangeability - and at an economical cost."

Thus the world at large has come to define quality not in absolute terms but rather in the context of the expectations of the customer, the purpose for which a product or service is acquired and how well it fits that purpose.

8.2 The Concept of Quality in Survey Research

Deming is mostly identified in the popular mind with the so-called "quality revolution" and the TQM movement, first in postwar Japan and then here in the US. Within the survey profession he is more likely thought of first and foremost as an eminent statistician who in the early 1940s was instrumental in the development of sampling procedures at the Bureau of the Census (Aguayo 1990). It was during his time at Census that Deming (1944) argued that accuracy should not be the sole criteria for evaluating survey data. It is at least equally important that survey results be "useful" by which he meant, "helping to provide a rational basis for action." (p. 369).

The link between quality and usefulness has been periodically reinforced throughout the literature on sampling and surveys. For example, Sudman (1976) tied sample quality to how the information will be used by describing two extremes. One of those extremes he calls "exploratory data gathering" whose main purpose is to generate initial hypotheses to be used in a later study. At the other extreme are large-scale government data collections to support policy development and program implementation. For Sudman the needs of the latter dictate a higher level of precision than the former. Thus, he seemed to adhere to a fitness for use test in which one important measure is the accuracy of the estimate. He offered 10 specific examples of studies in which sample designs were compromised to one degree or another, mostly due to cost or general feasibility. The emphasis,

however, in virtually all cases is on whether the potential loss of accuracy that flows from those compromises will substantially affect the usefulness of the estimates.

Kish (1987), in a similar vein, acknowledged that "statistical designs always involve compromises between the desirable and the possible." (p.1) For Kish those compromises are driven by the practicalities of feasibility and resources while being attuned to the purpose for which the research is designed. He listed three main categories or dimensions in which compromises typically are made: (1) *representation* having to do mostly with incomplete sampling frames and low nonresponse; (2) *randomization* meaning finding ways to account for the effects of confounding variables; and (3) *realism* which has to do with the degree to which survey variables measure the constructs they are meant to describe. He presented 10 research designs in an ordered list. At one end of the list were surveys where representation is essential. At the other end were experiments where randomization is critical. He placed observational studies, where the need for realism drives compromise, in the middle. This list of designs is meant to capture the differences between what Kish calls "enumerative" or descriptive studies on the one hand and "experimental" or analytical studies on the other.

Groves (1989) made a similar point when he noted how the needs of what he called "describers" differ from those of "modelers." The former look for survey data that fully reflect the population they want to describe and are especially concerned with errors of non-observation. The latter seek data with measures that fully capture the concepts needed to test their theories and worry less about coverage error and nonresponse. Government agencies are mostly describers because their principal interest is in a precise estimate of some specific characteristic(s) of a target population. Academics and sometimes market researchers, on the other hand, tend to be modelers who are interested in how personal characteristics interact to produce a specific behavior such as voting for one candidate over another or choosing product A rather than product B. Describers and modelers each have their own primary areas of concern and therefore different ideas about what constitutes high quality data.

O'Muircheartaigh (1997) also linked the needs of the data user and the reasons for which the data were collected to quality.

> The concept of quality, and indeed the concept of error, can only be defined satisfactorily in the same context as that in which the work is conducted. To the extent that the context varies, and the objectives vary, the meaning of error will also vary. . . Rather than specify an arbitrary (pseudo-objective) criterion, this redefines the problem in terms of the aims and frame of reference of the researcher. It immediately removes the need to consider true value concepts in any absolute sense, and forces consideration of the needs for which the data are being collected. (p.1).

8.3 Fit for purpose in Government Statistical Agencies

Government statistical agencies are the classic example of describers for whom accuracy is the primary attribute of quality, often because the estimates typically play a major role in making financial and policy decisions. Yet increasingly, a considerable body of statistical agency requirements includes additional quality dimensions and criteria by which fit for purpose design decisions might be made. In 2002, Statistics Canada released a set of quality guidelines that specified six "elements of quality . . . to be considered and balanced in the design and implementation of the agency's statistical programs." (p.4) The six are:

- *Relevance* –the degree to which the data are needed for some agency or policy purpose.
- *Accuracy* – the degree to which estimates correctly measure what they are designed to measure within an acceptable margin of error.
- *Timeliness* – the likelihood that the data will be available when it is needed to support an intended action.
- *Accessibility* – the availability of the data in a form that is useful to those needing them.
- *Interpretability* – the ease with which users of the data can understand the design and data collection processes so that judgments can be made about its usefulness.
- *Coherence* – the degree to which the data fit within a statistical program and are consistent with other similar data across time and geography.

The Australian Bureau of Statistics (2009) has developed a similar framework. It adds a seventh element of quality that it calls "the Institutional Environment." This refers to the brand of the

agency or organization associated with a statistical product and the degree to which it engenders confidence by virtue of a history of objectivity, independence, and protection of confidentiality of research participants.

Other agencies have frameworks that are variations on this same set of themes. They include Eurostat (2003), Statistics Sweden (Rosén and Evers 1999), the International Monetary Fund (Carson 2001), and the Organization for Economic Coordination and Development (OECD 2002).

8.4 Fit for purpose in Market Research

Market researchers sometimes behave as describers and other times as modelers. Tracking studies that continually measure phenomena such as product satisfaction or use over time are in some ways similar to data collections by government statistical agencies. Media measurement services that rack viewership and readership are another example of where precise estimates are desired. In these instances the fitness for use criteria applied are not unlike those described above for government agencies although an emphasis on cost, timeliness and accessibility sometimes predominate. At other times market researchers are more like modelers whose main focus is on data collections that can support development and testing of statistical models that describe, for instance, how personal characteristics and product features interact to make some products successful while others fail.

So market researchers, whether implicitly or explicitly, have adopted the concept of fitness for use. Nowhere is this clearer than in the specific context of opt-in panels (e.g., Bain, 2011). However, relatively little has been written to describe a specific framework of the sort used by the government statistical agencies. ESOMAR, the global professional organization for market, opinion, and social research, regularly issues guidelines to help researchers make fit to purpose decisions about specific research methodologies. For example, their "26 Questions to Help Research Buyers of Online Samples" (ESOMAR, 2008) suggests that a prospective online sample provider be vetted on seven criteria: the company's profile, its sample sources, recruitment methods, panel management practices, compliance with industry policy and applicable law, partnerships with other suppliers, and methods

of data quality and validation. Their "36 Questions to Help Commission Neuroscience Research" (ESOMAR, 2011) use a somewhat different set of criteria. As their titles suggest, both documents suggest questions to be asked, describe why each question may be important, but leave it to the researcher to evaluate how the answers may affect the quality of a specific research design.

Smith and Fletcher (2004) have offered a much clearer framework. They started from the premise that "there is an often unavoidable gap between the quality of the raw data and the kind of market research evidence to which we aspire" (p. 61). The challenge that falls primarily to the analyst is to bridge the gap between the ideal and the real, between the data we wish we had and the data that we have. To that end Smith and Fletcher offered eight methodological questions for analysts to ask, most of which might be interpreted in terms of the error categories in the standard TSE model. One of the eight involves a judgment about whether the study design was fit to purpose, which they characterized as a tradeoff among five key variables:

- The required accuracy/precision
- The depth or level of detail in the data
- The practical and ethical constraints
- The time when the data are needed
- The budget

8.5 Summary

Compromises in survey design are commonplace in surveys of all kinds and in all sectors of the research industry. Advocates of probability samples have come to accept what once they may have thought of as unacceptably high levels of nonresponse. The dramatic rise in the use of opt-in panels has been premised on a willingness to accept overwhelming coverage and selection error. Those compromises are mostly practical and increasingly accepted, but seldom explicitly set in a fitness for use framework. Yet even in idealized circumstances where the classic constraints of budget, time, and feasibility may not require compromise in design the degree to which survey results are usable by

decision makers is now widely recognized as a key driver of design. To quote Groves and Lyberg (2010), ". . . statistics are of little importance without their use being specified." (p. 873).

The notion of fit for purpose as a definition of quality and the key driver of survey design has some history. In their brief review of that history Biemer and Lyberg (2003) arrived at fitness for use as the basic definition of survey quality. They defined the three key dimensions of survey quality as accuracy, timeliness, and accessibility. In their words, for survey data to be considered of high quality it must be "as *accurate* as necessary to achieve their intended purposes, be available at the time it is needed (*timely*) and be *accessible* to those for whom the survey was conducted." (p.13). A number of government statistical agencies have further elaborated that basic framework, although it is our impression that these organizations have struggled with implementation.

Arguably the most compelling framework is the distinction between the needs and purposes of describers versus those of modelers. The needs of the former fit rather neatly into the TSE framework that has been widely adopted by survey researchers. Describers especially focus on minimizing coverage error and nonresponse as a way of ensuring representation and accuracy. They generally favor probability-based sample designs.

Modelers, on the other hand, are much more focused on measuring all of the concepts that they expect play a significant role in explaining the behavior that is at the core of their research. Their interest is in the relationships among a broad set of characteristics rather than precise measurement of those characteristics in the population of interest. They don't ignore errors of non-observation, but they generally assume no meaningful relationship among the phenomenon they are studying, the dependent variable(s), and the likelihood of an individual being selected or responding. Put another way, their primary emphasis tends to be on internal validity. Modelers make extensive use of non-probability samples because they often are less expensive than those favored by describers.

As useful as these two distinctions might be they are less a dichotomy than opposite ends of a continuum. Kish's ordered list of designs is probably the best framework for thinking about how we

might accommodate the purpose of the research and the precision required to the practicalities of time, budget, accessibility and general feasibility.

In some cases, the choice may come down to a non-probability sample or no survey at all. If some level of confidence that the assumptions of the model hold sufficiently for the purposes of the research, then the choice of a non-probability sample is justified. If not, the estimates from a non-probability sample might lead to poor decisions and no survey may be a better choice.

9. CONCLUSIONS

We believe that the foregoing review highlights the major issues survey researchers face when considering non-probability sampling. Due to the unsettled landscape that is non-probability sampling for surveys, our treatment of these issues may not be as definitive as would be possible in more developed areas of survey research. Many of the methods we described have long been used in other disciplines, but they remain unfamiliar to many in the survey profession. We hope that this report is the beginning of a much broader exploration of those methods and, to that end, summarize below what we believe to be the key issues.

Unlike probability sampling, there is no single framework that adequately encompasses all of non-probability sampling. Although there are variants within probability sampling, the design-based approach has a common theoretical basis – select a relatively large sample where the chance of selecting each unit is known and weight the sample by the inverse of its selection probability (adjusting for missing data as needed) to estimate population characteristics. By comparison, there are a variety of non-probability sampling methods, each with different approaches to sampling and estimation. The statistical properties and empirical performance of these methods vary considerably. Thus, non-probability sampling is a collection of methods rather than a single method, and it is difficult if not impossible to ascribe properties that apply to all non-probability sampling methodologies. Before using a non-probability method, a researcher should understand the specific approach being used and its limitations, especially with regard to limitations on the inferences that can be made.

Researchers and other data users may find it useful to think of the different non-probability sample approaches as falling on a continuum of expected accuracy of the estimates. At one end are uncontrolled convenience samples that produce estimates assuming that respondents are a random sample of the population. These surveys generally have little or no evidence to support them. Inferences from these samples are extremely risky. At the other end are

methods that select respondents based on criteria related to the survey subject matter and adjust the results using variables that are correlated with the key outcome variables. When studies also have evidence to support their assumptions, inferences from these samples are less risky. [3]

The difficulty arises in placing surveys between the two extremes. This is largely uncharted territory for social, opinion and market research surveys. The risk assessment often depends on both substantive knowledge of the population being studied and technical features of the methods being used. Substantively, understanding the variability of the characteristic(s) being studied within the target population and the planned use of the estimates are critical to assessing the risk of the inference. The most common error in studies of human behavior and attitudes is to assume more homogeneity than exists. This error leads to biases and understatements of the variability of the estimates. Technically, sampling and adjustment control may be even harder to evaluate. We believe that some control at sample selection is very important; probability sampling is an example of a highly controlled sampling mechanism. Rubin (2008) expresses a similar sentiment in another setting. The importance of controlling the sample is at the core of another conclusion (below) that sample matching is the most promising non-probability approach for surveys. Combining control of the sampling with the use of good auxiliary variables at the adjustment stage should make inferences from non-probability samples less risky.

Transparency is essential. Whenever non-probability sampling methods are used, there is a higher burden than that carried by probability samples to describe the methods used to draw the sample, collect the data, and make inferences. Once again, non-probability sampling is not a single method and differences in methods may be very important. Therefore, a clear description of methods and assumptions is essential for understanding the usefulness of the estimates. When the assumptions are clearly identified and there is evidence to support those assumptions, then the user

[3] In practice, few contemporary surveys, not even probability samples, might be classified as low risk using these criteria

has the opportunity to make an informed assessment of the risks associated with their use of the survey estimates. That assessment is not simple and it is a hurdle that non-probability samples must clear if they are to be generally accepted in scientific circles. Black box methodologies must be opened up and made transparent. Genuine proprietary information may not have to be disclosed, but the methodology must be clear and the effort made to assess key assumptions summarized. Too many online surveys, in particular, consistently fail to include information that is adequate to assess their methodology. This must change.

Making inferences for any probability or non-probability survey requires some reliance on modeling assumptions. Those assumptions must be clearly stated and evidence of the effect that departures from those assumptions might have on the accuracy of the estimates identified to the extent possible. While our focus has been on non-probability sampling methods, users of probability samples should also recognize that the inability to obtain 100% coverage and response rates means that inferences from these samples also often rely on modeling assumptions.

The most promising non-probability methods for surveys are those that are based on models that attempt to deal with challenges to inference in both the sampling and estimation stages. Although probability sampling is the standard paradigm for making statistical inferences from surveys, it is not the only way to make inferences or even the most common one in general statistical practice. Other approaches, called model-based when applied in the survey field (Valliant, Royall, and Dorfmann 2000), are used in most non-survey applications. These approaches typically assume that responses are generated according to a statistical model (e.g., the observations come from a process that has a common mean and variance), and that by accounting for important auxiliary variables it is possible to improve the fit and usability of these models. Once the model is formulated, standard statistical estimation procedures such as likelihood-based or Bayesian techniques are used to make inferences about the parameters being estimated. Of course, simple

model assumptions that ignore the complexity of the issues may not be sufficient. Even with model-based methods the challenges to inference remain.

One of the reasons model-based methods are not used more frequently in surveys may be that developing the appropriate models and testing their assumptions is difficult and time-consuming, requiring significant statistical expertise. Assumptions should be evaluated for all the key estimates, and a model that works well for some estimates may not work well for others. One of the key attributes of probability sampling is that it has a standard approach to produce the vast array of estimates often needed from surveys. Although probability sampling requires assumptions to deal with missing data, the assumptions are relatively standard and the approaches to adjustment have become routine. Achieving the simplicity of probability sampling methods for producing multiple estimates is a hurdle for non-probability sampling methods to overcome. In essence, even though non-probability samples can collect data much more cheaply than probability samples in many circumstances, the less controlled sampling approach requires more effort at the analysis stage.

Fit for purpose is an important concept for judging survey data quality, but its application to survey design requires further elaboration. In discussing complex modeling, a report of the National Academy of Science (2012) wrote, "the level of rigor employed should be commensurate with the importance and needs of the application and decision context. Some applications involve high-consequence decisions and therefore require a substantial …effort; others do not" (p. 96). There is an emerging consensus among national statistical agencies around the key elements that should make up a quality framework to support the development and execution of survey designs that are fit to purpose. They generally involve balancing the key elements of relevance, accuracy, timeliness, accessibility, interpretability, and consistency. The logical next step is to transition those frameworks from the theoretical to the practical, refining them accordingly. Non-probability sampling needs to be evaluated within this same framework.

Sampling methods used with opt-in panels have evolved significantly over time, and, as a result, research aimed at evaluating the validity of survey estimates from these sample sources should focus on sampling methods rather than the panels themselves. There is a tendency to lump all online samples from opt-in panels into the single bucket of online, as if opt-in panels were a sampling method. They are not. Users of opt-in panels may employ different sampling, data collection, and adjustment techniques. Research evaluations of older methods of non-probability sampling from panels may have little relevance to the current methods being used. Research should shift from a focus on opt-in panels to one that evaluates the different sampling and estimation strategies researchers might employ with these sample sources.

If non-probability samples are to gain wider acceptance among survey researchers there must be a more coherent framework and accompanying set of measures for evaluating their quality. One of the key advantages of probability sampling is the toolkit of measures and constructs (such as TSE) developed for it that provides ways of thinking about quality and error sources. Using that toolkit to evaluate non-probability samples is not especially helpful because the framework for sampling is different. Arguably the most pressing need is for researched aimed at developing better measures of the quality of non-probability sampling estimates that include bias and precision.

Although non-probability samples often have performed well in electoral polling, the evidence of their accuracy is less clear in other domains and in more complex surveys that measure many different phenomena. Surveys designed to yield only a handful of estimates on a related set of outcomes may require the control of only a small set of covariates. Frequent repetition, as in tracking surveys, and the availability of external benchmarks (such as election results) make experimentation possible and may make model development easier. However, many surveys do not have these advantages. A survey often produces many estimates across a broad array of subject areas and domains, requiring a larger set of covariates.

Non-probability samples may be appropriate for making statistical inferences, but the validity of the inferences rests on the appropriateness of the assumptions underlying the model and how deviations from those assumptions affect the specific estimates. Throughout the report, we have emphasized the need for further development of a theoretical basis for any non-probability sampling method to be followed by empirical evaluation of that method. The evaluation should assess the appropriateness of the assumptions under various circumstances and for different estimates. Our review identified sample matching as one of method that already has a theoretical basis constructed for evaluation studies that could be modified and amplified for use with surveys. Several researchers have begun this effort already. The post-survey adjustment methods applied to non-probability sampling have largely mirrored efforts in probability samples. Although this may be appropriate and effective to some extent, further consideration of selection bias mechanisms may be needed. We believe an agenda for advancing a method must include these attributes.

REFERENCES

AAPOR (American Association for Public Opinion Research). 2012. "Understanding a 'Credibility Interval' and How it Differs from the 'Margin of Sampling Error' in a Public Opinion Poll." Downloaded from http://aapor.org/AM/Template.cfm?Section=Understanding_a_credibility_interval_and_how_it_differs_from_the_margin_of_sampling_error_in_a_publi&Template=/CM/ContentDisplay.cfm&ContentID=5475

AAPOR. (American Association for Public Opinion Research). 2009. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Revised 2011.

Abate, Tom. 1998. "Accuracy of Online Surveys May Make Phone Polls Obsolete." *The San Francisco Chronicle*, D1.

Alvarez, R. Michael, and Carla VanBeselaere. 2005. "Web-Based Surveys." *The Encyclopedia of Measurement*. California Institute of Technology. http://www.mta.ca/~cvanbese/encyclopedia_new2.pdf.

Asur, Sitaram, and Bernardo A. Huberman. 2010. "Predicting the Future with Social Media." Downloaded from http://arxiv.org/pdf/1003.5699v1.

Australian Bureau of Statistics 2009. ABS Data Quality Framework, Downloaded from http://www.abs.gov.au/ausstats/abs@.nsf/mf/1520.0 on 4/30/2013.

Baker, Reg, Stephen J. Blumberg, J. Michael Brick, Mick P. Couper, Melanie Courtright, J. Michael Dennis, Don Dillman, Martin R. Frankel, Philip Garland, Robert M. Groves, Courtney Kennedy, Jon Krosnick, Paul J. Lavrakas, Sunghee Lee, Michael Link, Linda Piekarski, Kumer Rao, Randall K. Thomas, and Dan Zahs. 2010. "AAPOR Report on Online Panels." *Public Opinion Quarterly* 74(4):711–81.

Banks, David. 2011. "Reproducible Research: A Range of Response Statistics." *Politics, and Policy*, 2, DOI: 10.2202/2151–7509.1023.

Berinsky, Adam J. 2006. "American Public Opinion in the 1930s and 1940s: The Analysis Of Quota-Controlled Sample Survey Data." *Public Opinion Quarterly* 70(4):499–529.

Bernhardt, Annette, Ruth Milkman, Nik Theodore, Douglas Heckathorn, Mirabai Auer, James DeFilippis, Ana Luz González, Victor Narro, Jason Perelshteyn, Diane Polson, and Michael Spiller. 2009. "Broken Laws, Unprotected Workers: Violations of Employment and Labor Laws in America's Cities." Downloaded from http://www.nelp.org/page/-/brokenlaws/BrokenLawsReport2009.pdf?nocdn=1.

Berzofsky, Marcus E., Rick L. Williams, and Paul P. Biemer. 2009. "Combining Probability and Non-Probability Sampling Methods: Model-Aided Sampling and the O*NET Data Collection Program." *Survey Practice* August: 1-5. http://surveypractice.files.wordpress.com/2009/08/berzofsky.pdf.

Bethlehem, Jelke. 2010. "Selection Bias in Web Surveys." *International Statistical Review* 78(2):161–88.

Bethlehem, Jelke, and Silvia Biffignandi. 2012. *Handbook of Web Surveys*. Hoboken, New Jersey: John Wiley & Sons Inc.

Bethlehem, Jelke, Fannie Cobben, and Barry Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. Hoboken, NJ: John Wiley & Sons.

Biemer, Paul P., and Lars E. Lyberg. 2003. *Introduction to Survey Quality*. New York: Wiley.

Biernacki, Patrick, and Dan Waldorf. 1981. "Snowball Sampling: Problem And Techniques Of Chain Referral Sampling." *Sociological Methods and Research* 10(2):141–63.

Birnbaum, Zygmunt William, and Monroe G. Sirken. 1965. "Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates." National Center for Health Statistics. *Vital and Health Statistics* 2(11).

Blair, Johnny, and Frederick Conrad. 2011. "Sample Size for Cognitive Interview Pretesting." *Public Opinion Quarterly* 75(4):636–58.

Brick, J. Michael. 1990. "Multiplicity Sampling in an RDD Telephone Survey." Sampling Design Issues section of the *American Statistical Association* 296–301.

Brick, J. Michael. 2011. "The Future Of Survey Sampling." *Public Opinion Quarterly* 75(5):872–88.

Brick, J. Michael, and Douglas Williams. 2013. "Explaining Rising Nonresponse Rates in Cross-Sectional." *The ANNALS of the American Academy of Political and Social Science* 645(1):36–59.

Bryson, Maurice C. 1976. "The Literary Digest Poll: Making of a Statistical Myth." *American Statistical Association* 30(4):184–85.

Callegaro, Mario, and Charles DeSogra. 2008. "Computing Response Metrics for Online Panels." *Public Opinion Quarterly* 72(5):1008–32.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental And Quasi-Experimental Designs For Research*. Chicago, IL: Rand-McNally.

Carlson, Robert G., Jichuan Wang, Harvey A. Siegal, Russel S. Falck, Jie Guo. 1994. "An Ethnographic Approach To Targeted Sampling: Problems And Solutions In AIDS Prevention Research Among Injection Drug And Crack-Cocaine Users." *Human Organization* 53:279–86.

Centers for Disease Control and Prevention. 2005. "Statistical Methodology of the National Immunization Survey 1994-2002." *Vital and Health Statistics* Series 2, 138.

Chang, Linchiat, and Jon A. Krosnick. 2009. "National Surveys via RDD Telephone Interviewing Versus the Internet: Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 73(4):641–78.

Chui. Michael, Markus Löffler, and Roger Roberts. 2010. "The Internet of Things." *McKinsey Quarterly*.

Clinton, Joshua D. 2001. "Panel Bias from Attrition and Conditioning: A Case Study of the Knowledge Networks Panel." Paper presented at 56th Annual Conference of the American Association for Public Opinion Research, May, Montreal, Canada.

Cochran, William G. 1965. "The Planning of Observational Studies of Human Populations." JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES A, 128(2):234–66.

Coleman, James S., 1958. "Relational Analysis: The Study of Social Organizations with Survey Methods." *Human Organization* 17:28–36.

Copas, John B. and H. G. Li. 1997. "Inference for Non-random Samples." *Journal of the Royal Statistical Society* Series B, 59:5-95.

Cornfield, Jerome. 1971. "The University Group Diabetes Program: A Further Statistical Analysis of the Mortality Findings." *Journal of the American Medical Association* 217(12):1676–87.

Couper, Mick P., Arie Kapteyn, Matthias Schonlau, and Joachim Winter. 2007. "Noncoverage and Nonresponse in an Internet Survey." *Social Science Research* 36:131–48.

Couper, Mick P. 2000. "Web Surveys: A Review of Issues and Approaches." *Public Opinion Quarterly* 64(4):464–94.

Couper, Mick P. 2007. "Issues of Representation in eHealth Research with a Focus on Web Surveys." *American Journal of Preventive Medicine* 32:S83–S89.

Couper, Mick P., and Michael Bosnjak. 2010. "Internet Surveys." *Handbook of Survey Research* Chapter 16, P.V. Marsden and J.D. Wright editors, Bingley, UK: Emerald Group Publishing Limited.

Curtin, Richard, Stanley Presser, and Eleanor Singer. 2005. "Changes in Telephone Survey Nonresponse Over the Past Quarter Century." *Public Opinion Quarterly* 69(1):87–98.

Dawber, Thomas R., Gilcin F. Meadors, and Felix E. Moore. 1951. "Epidemiological Approaches to Heart Disease: The Framingham Study." *American Journal of Public Health* 41:279–86.

de Munnik, Daniel, David Dupuis, and Mark Illing. 2009. "Computing the Accuracy of Complex Non-Random Sampling Methods: The Case of the Bank of Canada's Business Outlook Survey." Bank of Canada Working Paper 2009–10, March 2009. http://www.bankofcanada.ca/wp-content/uploads/2010/02/wp09-10.pdf.

Dever, Jill A., Ann Rafferty, and Richard Valliant. 2008. "Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?" *Survey Research Methods* 2:47–62.

Deville, J.C., and Särndal, C.E. (1992). "Calibration estimators in survey sampling." *Journal of the American Statistical Association*, 87, 376-382.

Deville, J.C.. 1991. "A Theory of Quota Surveys." *Survey Methodology* 17:163–81.

Diamond, Shari S. 2000. "Reference Guide on Survey Research." In *Reference Manual on Scientific Evidence* 2nd Edition. Washington, DC: Federal Judicial Center.

DiSogra, Charles. 2008. "River Samples: A Good Catch for Researchers?" In GfK Knowledge Networks http://www.knowledgenetworks.com/accuracy/fall-winter2008/disogra.html

Duffield, Nick. 2004. "Sampling for Passive Internet Measurement: A Review." *Statistical Science* 19(3):472–498.

Duffy, Bobby, Kate Smith, George Terhanian, and John Bremer. 2005. "Comparing Data from Online and Face-to-Face Surveys." *International Journal of Market Research* 47:615–39.

Duncan, G. 2008. "When to Promote, and When to Avoid, a Population Perspective." *Demography* 45(4):763–84.

Efron, Brad and Rob Tibshirani. 1993. *An Introduction to the Bootstrap*. CRC Press.

Elliott, Marc, and Amelia Haviland. 2007. "Use of a Web-Based Convenience Sample to Supplement a Probability Sample." *Survey Methodology* 33(2):211–15.

Elliott, Michael R. 2009. "Combining Data from Probability and Non-Probability Samples Using Pseudo-Weights." *Survey Practice.* August. http://surveypractice.files.wordpress.com/2009/08/elliott.pdf.

Elżbieta, Getka-Wilczyńska. 2009. "Mathematical Modeling of the Internet Survey." Warsaw School of Economics, Poland. http://www.intechopen.com/source/pdfs/8946/InTech-Mathematical_modeling_of_the_internet_survey.pdf.

Erikson, Robert S. and Christopher Wlezien. 2008. "Are Political Markets Really Superior to Polls as Election Predictors?" *Public Opinion Quarterly* 72(2):190–215.

Eysenbach, Gunther. 2004. "Improving the Quality of Web Surveys: The Checklist for Reporting Results from Internet E-Surveys (CHERRIES)." *Journal of Medical Internet Research* 6(3):e34.

Fan, David P. 2011. "Representative Responses from Non-Representative Survey Samples." Paper presented at the Midwest Association of Public Opinion Research, Chicago.

Field, Lucy, Rachel A. Pruchno, Jennifer Bewley, Edward P. Lemay Jr, Norman G. Levinsky. 2006. "Using Probability vs. Nonprobability Sampling to Identify Hard-to-Access Participants for Health-Related Research: Costs and Contrasts." *Journal of Aging and Health* 18(4):565–83.

Felix-Medina, Martin H., and Steven K. Thompson. 2004. "Combining Link-Tracing Sampling and Cluster Sampling to Estimate the Size of Hidden Populations." *Journal of Official Statistics* 20(1):19–38.

Frank, Ove. 1971. *Statistical Inference in Graphs*. Ph.D. thesis, Stockholm.

Frank, Ove. 1995. "Network Sampling and Model Fitting." [Carrington, Peter J., John Scott, and Stanley Wasserman, eds.]. *Models and Methods in Social Network Analysis*. Cambridge University Press, 2005, 31–56. Cambridge Books Online. http://dx.doi.org/10.1017/CBO9780511811395.003.

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

Fricker, Scott, and Roger Tourangeau. 2010. "Examining the Relationship Between Nonresponse Propensity and Data Quality in Two National Household Surveys." *Public Opinion Quarterly* 74(5):934–55.

Fuller, Wayne A. 1987. *Measurement Error Models*. New York, NY: John Wiley & Sons Inc.

Gile, Krista J. 2011. "Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation." *Journal of the American Statistical Association* 106:135–46.

Gile, Krista J., and Mark S. Handcock. 2010. "Respondent-Driven Sampling: An Assessment of Current Methodology." *Sociological Methodology* 40:285–327.

Gile, Krista J., and Mark S. Handcock. 2011. "Network Model-Assisted Inference from Respondent-Driven Sampling Data." ArXiv Preprint.

Gile, Krista J., Lisa G. Johnston, and Matthew J. Salganik. 2012. "Diagnostics for Respondent-Driven Sampling." arXiv:1209.6254. Under Review.

Gini, Corraco, and Luigi Galvani. 1929. "Di una Applicazione del Metodo Representative." *Annali di Statistica* 6(4):1–107.

Gittelman, Steven H. and Elaine Trimarchi. 2009. "Consistency: The New Quality Concern." Marketing Research Association's *Alert! Magazine*, October, 49(10):19–21.

Gittleman, Steven H. and Elaine Trimarchi 2010. "Online Research…and All that Jazz! The Practical Adaptation of Old Tunes to Make New Music." Online Research 2010. Amsterdam: ESOMAR.

Gjoka, Minas, Maciej Kurant, Carter T. Butts, and Athina Markopoilou. 2011. "Practical Recommendations on Crawling Online Social Networks." *JSAC* special issue on Measurement of Internet Topologies 29(9):1872–92.

Glasser, Gerald J., and Gale D. Metzger. 1972. "Random-Digit Dialing as a Method of Telephone Sampling." *Journal of Marketing Research* 9:59–64.

Goel, Sharad, and Matthew J. Salganik. 2010. "Assessing Respondent-Driven Sampling." *Proceedings of the National Academy of Science of the United States of America* 107(15):6743–47.

Goel, Sharad, and Matthew J. Salganik. 2009. "Respondent-Driven Sampling as Markov Chain Monte Carlo." *Statistics in Medicine* 28(17):2202–29.

Goodman, Leo A. 1961. "Snowball Sampling." *Annals of Mathematical Statistics* 32:148–70.

Groves, Robert M. 1989. *Survey Errors and Survey Costs.* New York, NY: John Wiley & Sons Inc.

Groves, Robert M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70:646–75.

Groves, Robert M., and Lars Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74(5):849–79.

Groves, Robert M., Eleanor Singer, and Amy Corning. 2000. "Leverage-Saliency Theory of Survey Participation: Description and an Illustration." *Public Opinion Quarterly* 64:299–308.

Groves, Robert M., Floyd Fowler, Mick P. Couper, James Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*. Wiley: New York.

Groves, Robert M., Stanley Presser, and Sarah Dipko. 2004. "The Role of Topic Interest in Survey Participation Decisions." *Public Opinion Quarterly* 681:2–31.

Häder, S., and Gabler, S. 2003. "Sampling and Estimation." In *Cross-Cultural Survey Methods*, Janet A. Harkness, Fons J.R. van de Vijver, and Peter Ph. Mohler. (eds.). New York Wiley, pp. 117–34.

Handcock, Mark S., and Krista J. Gile. 2011. "On the Concept of Snowball Sampling." *Sociological Methodology* 41(1):367–71.

Hansen, Morris H., William N. Hurwitz. 1943. "On the Theory of Sampling from Finite Populations." *Annals of Mathematical Statistics* 14 (4):333–62.

Hansen, Morris H., William G. Madow, and Benjamin J. Tepping. 1983. "An Evaluation of Model Dependent and Probability-Sampling Inferences in Sample Surveys." *Journal of the American Statistical Association* 78(384):776–93.

Harris Interactive. 2004. "Final Pre-Election Harris Polls: Still Too Close to Call but Kerry Makes Modest Gains." The Harris Poll #87, November 2, 2004. http://www.harrisinteractive.com/harris_poll/index.asp?pid=515.

Harris Interactive. 2008. "Election Results Further Validate Efficacy of Harris Interactive's Online Methodology." Press Release from Harris Interactive, November 6, 2008.

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

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47:153–62. http://vanpelt.sonoma.edu/users/c/cuellar/econ411/heckman.pdf.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "Beyond WEIRD: Towards a Broad-Based Behavioral Science." *Behavioral and Brain Sciences* 33:111–35.

Intrade. 2012. http://www.intrade.com/v4/home/

Ipsos. 2012. "Ipsos Poll Conducted for Reuters." Downloaded from www.ipsos-na.com/download/pr.aspx?id=11637

Isaksson, Annica and Sunghee Lee. 2005. "Simple Approaches to Estimating the Variance of the Propensity Score Weighted Estimator Applied on Volunteer Panel Web Survey Data - A

Comparative Study." SRMS proceedings. 3143–3149 http://www.amstat.org/sections/srms/proceedings/y2005/Files/JSM2005-000172.pdf

Pasek, Josh, and Krosnick, Jon A. 2010. "Measuring Intent to Participate and Participation in the 2010 Census and Their Correlates and Trends: Comparisons of RDD Telephone and Non-Probability Sample Internet Survey Data." Census.gov Study Series: Survey Methodology #2010-15, Statistical Research Division, U.S. Census Bureau. https://www.census.gov/srd/papers/pdf/ssm2010-15.pdf

Kalton, Graham. 1993. *Sampling Rare and Elusive Populations*. Department of Economic and Social Information and Policy Analysis Statistics Division, United Nations. New York.

Kalton, Graham. 2003. "Practical Methods for Sampling Rare and Mobile Populations." *Statistics in Transition* 6:491–501.

Kalton, Graham. 2009. "Methods of Oversampling Rare Population in Social Surveys." *Survey Methodology* December 2009, 35(2):125–41.

Kalton, Graham, and Dallas W. Anderson. 1986. "Sampling Rare Populations." *Journal of the Royal Statistical Society* Series A, 149(1):65–82.

Kalton, Graham, and Ismael Flores-Cervantes. 2003. "Weighting Methods." *Journal of Official Statistics* 19(2):81–97.

Kaplan, Charles D., Dirk Korf, Claire Sterk. 1987. "Temporal and Social Contexts of Heroin-Using Populations: An Illustration of the Snowball Sampling Technique." *The Journal Nervous and Mental Disease* 175(9):566–74.

Kendall, Carl, Ligia Kerr, Rogerio Gondim, Guilherme Werneck, Raimunda Macena, Marta Pontes, Lisa Johnston, Keith Sabin, and Willi McFarland. 2008. "An Empirical Comparison of Respondent-Driven Sampling, Time Location Sampling, and Snowball Sampling for Behavioral Surveillance in Men Who Have Sex with Men, Fortaleza, Brazil." *AIDS and Behavior*, 12(1):97–104.

Kiaer, Anders Nicolai. 1895-6. "Observations et Experiences Concernant des Denobrements Represenatifs." *Bulletin of the International Statistical Institute* 9, Liv. 2:176–83.

Kind, Allison. 2012. Tweeting the News, Case Study: News Organizations' Twitter Coverage of the 2011 State of the Union Address. http://www.american.edu/soc/communication/upload/Allison-Kind.pdf

Kish, Leslie. 1987. *Statistical Design for Research*. John Wiley & Sons, New York.

Kish, Leslie. 1995. "The Hundred Years' War of Survey Sampling." *Statistics in Transition* 2(5):813–30. Reprinted in 2003, Graham Kalton and Steven Heeringa, eds. *Leslie Kish: Selected papers*. New York, Wiley.

Kish, Leslie. 1965. "Selection Techniques for Rare Traits." *Genetics and the Epidemiology of Chronic Diseases*, Public Health Service Publication No. 1163.

Kish, Leslie. 1965. *Survey Sampling*. New York: John Wiley & Sons, Inc.

Kleinbaum, David G., Hal Morgenstern, and Lawrence L. Kupper. 1981. "Selection Bias in Epidemiologic Studies." *American Journal of Epidemiology* 113(4):452–63.

Klovdahl, Alden S., John J. Potterat, Donald E. Woodhouse, John B. Muth, Stephen Q. Muth, and William W. Darrow. 1994. "Social Networks and Infectious Disease: The Colorado Springs Study." *Social Science & Medicine* 38(1):79–88.

Kogan, Steven M., Cyprian Wejnert, Yi-fu Chen, Gene H. Brody, and LaTrina M. Slater. 2011. "Respondent-Driven Sampling with Hard-to-Reach Emerging Adults: An Introduction and Case Study with Rural African Americans." *Journal of Adolescent Research* 26(1):30–60.

Kott, Phillip S. 2006. "Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors." *Survey Methodology* 32(2):133–142.

Kreuter, Frauke, Kristen Olson, James Wagner, Ting Yan, Trena M. Ezzati-Rice, Carolina Casas-Cordero, Michael Lemay, Andy Peytchev, Robert M. Groves, and Trivelore Raghunathan. 2011. "Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys." *Journal of the Royal Statistical Society* Series A 173(2):389–407.

Kruskal, William, and Frederick Mosteller. 1980. "Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939." INTERNATIONAL STATISTICAL REVIEW 48:169–95.

Kruskal, William, and Frederick Mosteller. 1981. "Ideas of Representative Sampling." NEW DIRECTIONS FOR METHODOLOGY OF SOCIAL AND BEHAVIORAL SCIENCE: PROBLEMS WITH LANGUAGE IMPRECISION 3–24.

Lagakos, Stephen W., and Louise M. Ryan. 1985. "On the Representativeness Assumption in Prevalence Tests of Carcinogenicity." APPLIED STATISTICS 34:54–62.

Lavrakas, Paul J., Charles D. Shuttles, Charlotte Steeh, and Howard Fienberg. 2007. "The State of Surveying Cell Phone Numbers in the United States–2007 Special Issue." *Public Opinion Quarterly* 71(5):840–854.

Lee, Sunghee. 2006. "An Evaluation of Nonresponse and Coverage Errors in a Web Panel Survey." *Social Science Computer Review* 2(4):460–75. http://ssc.sagepub.com/content/24/4/460.abstract.

Lee, Sunghee. 2006. "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys." *Journal of Official Statistics* 22(2):329–49.

Lee, Sunghee, and Richard Valliant. 2009. "Estimation for Volunteer Panel Web Surveys using Propensity Score Adjustment and Calibration Adjustment." *Sociological Methods and Research* 37(3):319-43. http://smr.sagepub.com/content/37/3/319.full.pdf.

Lee, Sunghee, and Richard Valliant. 2008. "Weighting Telephone Samples Using Propensity Scores." In *Advances in Telephone Survey Methodology,* edited by J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. de Leeuw, L. Japec, P. J. Lavrakas, M. W. Link, and R. L. Sangster, 170–83. Hoboken, NJ: John Wiley & Sons, Inc.

Lensvelt-Mulders, Gerty J. L. M., Peter J. Lugtig, and Marianne Hubregtse. 2009. September. "Separating Selection Bias and Non-Coverage in Internet Panels Using Propensity Matching." *Survey Practice,* 2 [e-journal].

Lesser, Virginia. "Advantages and Disadvantages of Probability and Non-Probability-Based Surveys of the Elderly and Disabled." [online presentation]. http://ncat.oregonstate.edu/pubs/TRANSED/1081_Surveys.pdf

Levy, Paul S., and Stanley Lemeshow. 2008. *Sampling of Populations: Methods and Applications.* 4th ed. Hoboken, NJ: John Wiley & Sons, Inc.

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

Little, Roderick J.A., and Sonia L. Vartivarian. 2004. "Does Weighting for Nonresponse Increase the Variance Of Survey Means?" April 2004. Working Paper 35. THE UNIVERSITY OF MICHIGAN DEPARTMENT OF BIOSTATISTICS WORKING PAPER SERIES.

Little, Roderick J.A. 1986. "Survey Nonresponse Adjustments for Estimates of Means." *International Statistical Review* 54(2):139–57.

Little, Roderick J.A. 1988. "Missing-Data Adjustments in Large Surveys." *Journal of Business & Economic Statistics* 63:287–296.

Lohr, Sharon L. 1999. *Sampling: Design and Analysis.* Pacific Grove, CA: Brooks/Cole Publishing Company.

Loosveldt, Geert, and Nathalie Sonck. 2008. "An Evaluation of the Weighting Procedures for an Online Access Panel Survey." *Survey Research Methods* 2:93–105.

Lynn, Peter. 2004. "The Use of Substitution in Surveys." *The Survey Statistician* 49:14–6.

Lynn, Peter, and Roger Jowell. 1996. "How Might Opinion Polls Be Improved? The Case for Probability Sampling." *Journal of the Royal Statistical Society* Series A 15:21–8.

MacKellar, Duncan, Kathleen M. Gallagher, Teresa Finlayson, Travis Sanchez, Amy Lansky, and Patrick S. Sullivan. 2007. "Surveillance of HIV Risk and Prevention Behaviors of Men Who Have Sex With Men—A National Application of Venue-Based, Time-Space Sampling." *Public Health Reports* 122(1):39–47.

Malhotra, Neil, and Jon A. Krosnick. 2007. "The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples." *Political Analysis* 15:286–323.

Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. 2011. *Big Data: The Next Frontier for Innovation, Competition, and Productivity.* Seoul, San Francisco, London, and Washington, DC: McKinsey Global Institute. http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation.

Matthews, Vince. 2008. "Probability or Nonprobability: A Survey is a Survey–or Is It?" National Agricultural Statistics Service (NASS) white paper. http://www.nass.usda.gov/Education_and_Outreach/Understanding_Statistics/Statistical_Aspects_of_Surveys/survey_is_survey.pdf

Mayes, Linda C., Ralph I. Horwitz, and Alvan R. Feinstein. 1988. "A Collection of 56 Topics with Contradictory Results in Case-Control Research." *International Journal of Epidemiology* 17:680–5.

McKenzie, David. J, and Johan Mistiaen. 2009. "Surveying Migrant Households: A Comparison of Census-Based, Snowball and Intercept Point Surveys." *Journal of the Royal Statistical Society* Series A, 172:339–60.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27:415–44.

Moser, Claus Adolf, and Alan Stuart. 1953. "An Experimental Study of Quota Sampling." *Journal of the Royal Statistical Society: Series A,* 116:349–405.

Mosteller, Frederick, Herbert Hyman, Philip J. McCarthy, Eli S. Marks, David B. Truman, et al. 1949. "The Pre-Election Polls of 1948." Report to the Committee on Analysis of Pre-Election Polls and Forecasts. Bulletin 60. New York: Social Science Research Council.

Muhib, Farzana B., Lillian S. Lin, Ann Stueve, Robin L. Miller, Wesley L. Ford, Wayne D. Johnson, and Philip J. Smith. 2001. "A Venue-Based Method for Sampling Hard-To-Reach Populations." *Public Health Reports* 116(1):216–22.

Mutz, Diana. 2011. *Population-Based Survey Experiments.* Princeton, NJ: Princeton University Press.

National Research Council. 2012. Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification. Washington, D.C.: The National Academies Press.

New York Times. 2008. "Google Uses Searches to Track Flu's Spread." November 11. http://www.nytimes.com/2008/11/12/technology/internet/12flu.html?_r=2.

New York Times. 2012. "Which Polls Fared Best and Worse in the 2012 Presidential Race." November 10. http://fivethirtyeight.blogs.nytimes.com/2012/11/10/which-polls-fared-best-and-worst-in-the-2012-presidential-race/.

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.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." In *Proceedings of the Fourth International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media,* Washington, DC, May 2010. www.aaai.org.

Office of Management and Budget (OMB). 2006. *Standards and Guidelines for Statistical Surveys.* http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf. (DUPLICATE—SEE U.S. OFFICE OF MANAGEMENT AND BUDGET page 7)

Olivier, Lex. 2011. "River Sampling Non-Probability Sampling in an Online Environment." [Web log, November 13, 2011.] Center for Information-Based Decision Making and Marketing Research. http://lexolivier.blogspot.com/2011/11/river-sampling-non-probability-sampling.html

Olson, Kristen. 2006. "Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias." *Public Opinion Quarterly* 70(5):737–58.

Peracchi, Franco, and Finis Welch. 1995. "How Representative are Matched Cross-Sections? Evidence from the Current Population Survey." JOURNAL OF ECONOMETRICS 68:153–79.

Pew Research Center. 2011. "Muslim Americans: No Signs of Growth in Alienation or Support for Extremism." Survey Report. Washington, DC: Pew Research Center. http://www.people-press.org/2011/08/30/muslim-americans-no-signs-of-growth-in-alienation-or-support-for-extremism/

Peytchev, Andy, Sarah Riley, Jeffrey Rosen, Joe Murphy, and Mark Lindblad. 2010. "Reduction of Nonresponse Bias in Surveys Through Case Prioritization." *Survey Research Methods* 4(1):21–9.

Pfeffermann, Danny, and C.R. Rao, editors. 2009. *Sample Surveys: Inference and Analysis.* Handbook of Statistics, Vol 29B. Oxford: Elsevier B.V.

Poynter, Ray. 2010. *The Handbook of Online and Social Media Research.* Chichester, United Kingdom: Wiley.

Presser, Stanley. 1984. "The Use of Survey Data in Basic Research in the Social Sciences." In *Surveying Subjective Phenomena*, vol. 2, edited by C. F. Turner and E. Martin, 93–114. New York: Russell Sage.

Public Works and Government Services Canada. 2008. "The Advisory Panel on Online Public Opinion Survey Quality —Final Report." http://www.tpsgc-pwgsc.gc.ca/rop-por/rapports-reports/comiteenligne-panelonline/tdm-toc-eng.html

Rao, J.N.K. and C.F.J. Wu. 1988. "Resampling Inference with Complex Survey Data." *Journal of the American Statistical Association* 83:231–41.

Rao, John N.K. 2003. *Small Area Estimation.* Hoboken, NJ: John Wiley & Sons, Inc.

Ribeiro, Bruno, and Don Towsley. 2010. "Estimating and Sampling Graphs with Multidimensional Random Walks." In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement. http://arxiv.org/pdf/1002.1751.pdf.

Rivers, Douglas. 2007. "Sampling for Web Surveys." White paper prepared from presentation given at the 2007 Joint Statistical Meetings, Salt Lake City, Utah, July-August. https://s3.amazonaws.com/yg-public/Scientific/Sample+Matching_JSM.pdf.

Rivers, Douglas, and Delia Bailey. 2009. "Inference from Matched Samples in the 2008 U.S. National Elections." Paper presented at the 64th Annual Conference of the American Association for Public Opinion Research, Hollywood, Florida, May.

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

Robinson, William T., Jan M.H. Risser, Shanell McGoy, Adam B. Becker, Hafeez Rehman, Mary Jefferson, Vivian Griffin, Marcia Wolverton, and Stephanie Tortu. 2006. "Recruiting Injection Drug users: A Three-Site Comparison of Results and Experiences with Respondent-Driven and Targeted Sampling Procedures." *Journal of Urban Health* 83(1):29–38.

Rosenbaum, Paul R. 2005. "Observational Study." in Everitt & Howell, eds.l *Encyclopedia of Statistics in Behavioral Science*. Vol 3, 1451–1462.

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

Rosenbaum, Paul R., and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79:516–24.

Rothman, Kenneth J., and Sander Greenland. 1998. *Modern Epidemiology*. 2nd ed. Philadelphia: Lippincott Williams & Wilkins.

Rothman, Kenneth J., Sander Greenland, and Timothy L. Lash. 2008. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincottt Williams & Wilkins.

Rothschild, David. 2009. "Forecasting Elections: Comparing Prediction Markets, Polls, and Their Biases." *Public Opinion Quarterly* 73(5):895–916.

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

Rubin, Donald B. 2008. "For objective causal inference, design trumps analysis." *The Annals of Applied Statistics*, 2, 808-840.

Rubin, Donald B. 1979. "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association* 74:318–28.

Salganik, Matthew J. 2006. "Variance Estimation, Design Effects, and Sample Size Calculations for Respondent-Driven Sampling." *Journal of Urban Health* 83:98–112.

Salganik, Matthew J., and Douglas D. Heckathorn. 2004. "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling." *Sociological Methodology* 34:193–239.

Särndal, Carl-Erik, Bengt Swensson, and Jan Wretman. 1992. *Model-Assisted Survey Sampling*. New York: Springer-Verlag, Inc.

Savage, Mike, and Roger Burrows. 2007. "The coming crisis of empirical sociology." *Sociology,* 41, 885-899.

Schield, Milo. 1994. "Random Sampling Versus Representative Samples." *American Statistical Association Proceedings Of The Section On Statistical Education* 107–10. Downloaded from www.StatLit.org/pdf/1994SchieldASA.pdf.

Schillewaert, Niels, Tom De Ruyck, and Annelies Verhaeghe. 2009. "Connected Research – How Market Research Can Get the Most Out of Semantic Web Waves." *International Journal of Market Research* 51(1):11–27.

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

Schonlau, Matthias, Arthur van Soest, Arie Kapteyn, and Mick Couper. 2009. "Selection Bias in Web Surveys and the Use of Propensity Scores." *Sociological Methods & Research* 37:291–318.

Schonlau, Matthias, Kinga Zapert, Lisa Payne Simon, Katherine Sanstad, Sue Marcus, John Adams, Mark Spranca, Hongjun Kan, Rachel Turner, and Sandra Berry. 2004. "A Comparison Between Responses from a Propensity-Weighted Web Survey and an Identical RDD Survey." *Social Science Computer Review* 22:128–38.

Sears, David O. 1986. "College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51:515–30.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. 2nd ed. Stamford, CT: Cengage Learning/Wadsworth Publishing.

Shadish, William R., M. H. Clark, and Peter Steiner. 2008. "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments." *Journal of the American Statistical Association* 103:1334–44.

Silver, Nate. 2012. (See New York Times 2012 above.)

Smith, Aaron. 2011. *Trends in Cell Phone Usage and Ownership*. Washington, D.C.: Pew Research Center. http://www.pewinternet.org/Presentations/2011/Apr/FTC-Debt-Collection-Workshop-Cell-Phone-Trends.aspx

Smith, T. M. F. 1983. "On The Validity of Inferences From Non-Random Sample." *Journal of the Royal Statistical Society* Series A, 146(4):394–403.

Snell, Laurie, J., Peterson, Bill and Grinstead, Charles. (1998). "Chance News 7.11." Downloaded from http://www.dartmouth.edu/~chance/chance_news/recent_news/chance_news_7.11.html on August 31, 2009. Stirton Stirton

Snow, Rob E., John D. Hutcheson, James E. Prather. 1981. "Using Reputational Sampling to Identify Residential Clusters of Minorities Dispersed in a Large Urban Region: Hispanics in Atlanta, Georgia." Proceedings of the section on Survey Research Methods, *American Statistical Association* 101–6.

Squire, Peverill. 1988. "Why the 1936 Literary Digest Poll Failed." *Public Opinion Quarterly* 52:125–33.

Statistics Canada 2002. Statistics Canada's Quality Assurance Framework. Available from http://www5.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-586-X&CHROPG=1&lang=eng.

Statistics Canada. 2009. *Statistics: Power from Data! Nonprobability sampling.* http://www.statcan.gc.ca/edu/power-pouvoir/ch13/nonprob/5214898-eng.htm

Steckler, Allan, and Kenneth R. McLeroy. 2008. "The Importance of External Validity." *American Journal of Public Health* 98:9–10.

Steiner, Peter M., Thomas D. Cook, and William R. 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, Franklin F., and Philip J. McCarthy. 1958. *Sampling Opinions: An Analysis of Survey Procedure.* Oxford: John Wiley & Sons, Inc.

Stolzenberg, Ross M., and Daniel A. Relles. 1997. "Tools for Intuition About Sample Selection Bias and its Correction." *American Sociological Review* 62:494–507.

Stone, Mary Bishop, Joseph L. Lyon, Sara Ellis Simonsen, George L. White, and Stephen C. Alder. 2007. "An Internet-Based Method of Selecting Control Populations for Epidemiological Studies." *Practice of Epidemiology* 165:109–12.

Strauss, Murray A. 2009. "Validity of Cross-National Research Using Unrepresentative Convenience Samples." *Survey Practice* 43(3).

Sudman, Seymour. 1966. "Probability Sampling with Quotas." *Journal of the American Statistical Association* 20:749–71.

Sudman, Seymour, and Brian Wansink. 2002. *Building a Successful Convenience Panel.* Chicago, IL: American Marketing Association.

Sugden, R.A. and Smith, T.M.F. 1984. "Ignorable and Informative Designs in Survey Sampling Inference." *Biometrika* 71(3):495–506.

Taylor, Humphrey. 2007. "The Case for Publishing (Some) Online Polls." Polling Report.

Terhanian, George, and John Bremer. 2012. "A Smarter Way to Select Respondents for Surveys?" *International Journal of Market Research* 54(6):751–780.

Terhanian, George, and John Bremer. 2000. "Confronting the Selection-Bias and Learning Effects Problems Associated with Internet Research." Research paper: Harris Interactive. http://growingupwithmedia.com/pdf/Confronting_Selection_Bias.pdf.

Terhanian, George, and John Bremer. 1995. "Creative Applications of Selection Bias Modeling in Market Research." ISI paper.

Terhanian, George, Jonathan W. Siegel, Cary Overmeyer, John Bremer, and Humphrey Taylor. 2001. "The Record of Internet-Based Opinion Polls in Predicting the Results of 72 Races in the November 2000 U.S. Elections." *International Journal of Market Research* 43(2):127–135.

The Telegraph. 2012. "Wisdom Index Poll Puts Labour Eight Points Ahead of Conservatives." http://www.telegraph.co.uk/comment/9307396/Wisdom-index-poll-puts-Labour-eight-points-ahead-of-Conservatives.html

Thompson, Steven K. 1990. "Adaptive Cluster Sampling." *Journal of the American Statistical Association* 85:1050–59.

Thompson, Steven K. 1992. *Sampling.* Wiley, New York.

Thompson, Steven K., Linda M. Collins. 2002. "Adaptive Sampling in Research on Risk-Related Behaviors." *Drug and Alcohol Dependence* 68(1):S57–67.

Thompson, Steven K., and Ove Frank. 2000. "Model-Based Estimation with Link-Tracing Sampling Designs." *Survey Methodology* 26:87–98.

Thompson, Steven K., George A.F. Seber. 1996. *Adaptive Sampling.* Wiley, New York.

Thompson, Steven K. 2006a. "Targeted Random Walk Designs." *Survey Methodology* 32:11–24.

Thompson, Steven K. 2006b. "Adaptive Web Sampling." *Biometrics* 62:1224–34.

Thompson, Steven K. 1990. "Adaptive Cluster Sampling." *Journal of the American Statistical Association* 85(412):1050–59.

Thompson, Steven K. 2002. *Sampling*. New York: John Wiley & Sons, Inc.

Tighe, Elizabeth, David Livert, Melissa Barnett, and Leonard Saxe. 2010. "Cross-Survey Analysis to Estimate Low-Incidence Religious Groups." *Sociological Methods & Research* 39:56–82.

Tourangeau, Roger, Frederick G. Conrad, and Mick P. Couper. 2013. *The Science of Web Surveys*. New York: Oxford University Press.

Trow, Martin. 1957. *Right-Wing Radicalism and Political Intolerance*. Arno Press, New York, Reprinted 1980.

Twyman, Joe. 2008. "Getting It Right: Yougov and Online Survey Research In Britain." *Journal of Elections, Public Opinion & Parties* 18:343-354.

U.S. Census Bureau. 2011. *U.S. Census Bureau Statistical Quality Standards*. Washington, D.C.

U.S. Department of Agriculture (USDA). 2006. "The Yield Output Forecasting Program of NASS." Downloaded from http://www.nass.usda.gov/Education_and_Outreach/Understanding_Statistics/yldfrcst2006.pdf

U.S. Office of Management and Budget. 2006. *Standards and Guidelines for Statistical Surveys*. Washington, D.C. http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf

Valliant, Richard and Jill A. Dever. 2011. "Estimating Propensity Adjustments for Volunteer Web Surveys." *Sociological Methods & Research* 40(1):105–137. http://smr.sagepub.com/content/40/1/105.

Valliant, Richard, Alan H. Dorfman, and Richard M. Royall. 2000. *Finite Population Sampling and Inference*. New York: John Wiley & Sons, Inc.

Vavreck, Lynn and Rivers, Douglas. 2008. "The 2006 Cooperative Congressional Election Study." *Journal of Elections, Public Opinion & Parties* 18(4):355–66.

Vehovar, Vasja. 1995. "Field Substitutions in Slovene Public Opinion Survey." In *Contributions to Methodology and Statistics* Metodološki zvezki, 10. A. Ferligoj and A. Kramberger eds. Ljubljana: FDV, 39–66.

Vehovar, Vasja. 1999. "Field Substitution and Unit Nonresponse." *Journal of Official Statistics* 15(2):335–50.

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

Walker, Robert, and Raymond Pettit. 2009. *ARF Foundations of Quality: Results preview*. New York: The Advertising Research Foundation.

Watters, John K., and Patrick Biernacki. 1989. "Targeted Sampling: Options for the Study of Hidden Populations." *Social Problems* 36(4):416–30.

Wejnert, Cyprian, and Douglas D. Heckathorn. 2007. "Web-Based Network Sampling: Efficiency and Efficacy of Respondent-Driven Sampling for Online Research." *Sociological Methods and Research* 37(1):105–34.

Welch, Susan. 1975. "Sampling by Referral in a Dispersed Population." *Public Opinion Quarterly* 39:237–45.

Winship, Christopher, and Robert D. Mare. 1992. "Models for Sample Selection Bias." *Annual Review of Sociology* 18:327–350.

Wolter, Kirk M. 2007. *Introduction to Variance Estimation*. New York: Springer Science+Business Media, LLC.

Yates, F. 1946. "A Review of Recent Statistical Developments in Sampling and Sampling Surveys." *Journal of the Royal Statistical Society* 109:12–43.

Yeager, David S., Jon A. Krosnick, LinChiat Chang, Harold S. Javitz, Matthew S. Levendusky, Alberto Simpser, and Rui 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.

Zea, Maria Cecilia. 2010. "Reaction to the Special Issue on Centralizing the Experiences of LGB People of Color in Counseling Psychology." *Counseling Psychologist* 38(3):425–33.

Zukin, Cliff, Jessica Godofsky, Carl Van Horn, Wendy Mansfield, and J. Michael Dennis. 2011. "Can a Non-Probability Sample Ever Be Useful for Representing a Population?: Comparing Probability and Non-Probability Samples of Recent College Graduates." Research presented at the 2011 Annual Conference of the American Association for Public Opinion Research. http://www.knowledgenetworks.com/ganp/docs/aapor2011/aapor11-Can-a-Non-Probability-Sample.pdf

APPENDIX A: AAPOR NON-PROBABILITY TASK FORCE MISSION STATEMENT

November 22, 2011

Survey researchers routinely conduct studies that use different methods of data collection and inference. Some employ a probability sampling framework, while others use non-probability designs. There are a wide range of these non-probability designs including case-control studies, clinical trials, evaluation designs, intercept surveys, and volunteer panels. Researchers may also work with other datasets such as administrative records or surveillance systems in which the researcher does not have control of part or all of the system of collection. While not surveys per se, they nonetheless may provide insight into methods of working with data not collected using probability-based methods.

The mission of the task force is to examine the conditions under which various survey designs that do not use probability samples might still be useful for making inferences to a larger population. We assume that a large probability sample with complete coverage of the target population and complete response meets this requirement, but this rarely occurs in practice. Non-probability samples may be an acceptable alternative, depending on the purpose for which the data are being collected.

Non-probability sampling and inference designs are less well developed and explored. The assumptions required to make valid inferences from non-probability samples are not well known. Hence it is beneficial to examine non-probability methods and assumptions to assess the conditions under which they might be expected to provide estimates that are fit for the planned use.

The task force is charged with examining the strengths and weaknesses of different non-probability methods. Both theoretical and empirical support for these methods will be considered. Since online sampling, whether from access panels or intercepts, has become so prevalent, these designs will be a major but not exclusive focus of the task force.