Contents

1. Introduction
2. Random Sampling
   2.1. Standard Quota Sampling
   2.2. Quotas as a Set of Probabilities
3. Historical Antecedents of the “Parallel Surveys” Approach
   3.1. Marking a Fair Comparison
4. Inverting the Analytical Focus of Matching Methodologies
   4.1. The Harris Interactive Case Study
   4.2. The Harris Interactive Methodology
5. The Propensity Score Select Methodology
6. New Empirical Evidence
   6.1. Simulating the Effects of the Propensity Score Select Methodology
7. Assessing Accuracy
   7.1. “Demographic, Quota” Questions
   7.2. “Demographic (Non-Quota) and Household” Questions
   7.3. Overall Accuracy of All “Demographic and Household” Questions
   7.4. “General” Questions
   7.5. “Attitudes Towards Privacy” Questions
   7.6. “Technology Ownership” Questions
   7.7. “Online Behaviors” Questions
   7.8. Accuracy of All “Content” Questions
8. Next Steps and Closing Thoughts
9. References

Appendix. The Questions Used in the Survey
A Smarter Way to Select Respondents for Surveys

George H. Terhanian and John F. Bremer

26 February 2012

1. Introduction

Fifteen years ago, market researchers knew very little about online survey research. At the time, some were dabbling in the practice but nearly all interviews were still completed via telephone, paper-and-pencil, or face-to-face. Today, they know much more about online survey research and most interviews (specifically, those commissioned by US market research buyers) are completed online, with spending expected to exceed $1.8 billion in 2012, or half of all survey research spending (“Special Report: US Online MR Spend...,” 2012, p. 15). To keep up with online demand, many market research companies have developed their own opt-in panels (“Special Report: US Online MR Spend...,” 2012, p. 16). Some of these same companies have also launched new methods of accessing potential survey respondents, such as “river sampling” whereby online users, after clicking through an invitation or advertisement on a web site, are directed to a survey for which they might qualify. “Routing” systems have been introduced as well. Among other capabilities, they direct individuals who do not qualify for a survey to another, which can increase a type of utilization.

In some respects, however, online research has been a victim of its own success. Despite the increase in the number of panels and the development of new ways of accessing and utilizing respondents, researchers can still find it difficult to complete the needed number of interviews on time, particularly when the target population is rare or in high demand. For all of these reasons, it is common today for researchers to use more than one sample source for some types of surveys.

By reducing the severity of one problem, however, they may have created another. Adding one or more sample source to the original might address the need for more respondents, but some evidence suggests that it might also increase bias, defined here and elsewhere (at least roughly) as the difference between what survey respondents report and what we, or an omniscient observer, know to be true (Bohrnstedt, 1983). In research evaluating seventeen different opt-in panels in 2008, for instance, the Advertising Research Foundation found “wide variance, particularly on attitudinal and/or opinion questions (purchase intent, concept reaction, and the like),” even after holding constant socio-demographic and other factors (Walker, Pettit, and Rubinson, 2009).
To learn how to select multiple sample sources for the same survey in a way that reduces as much bias as possible, some new research has been undertaken to apparent good effect. For instance, Global Market Insights (GMI) has described its company’s Pinnacle methodology glowingly at several industry conferences (e.g., see Eggers, 2011).

Although we could not obtain a technical description of the Pinnacle methodology, we believe it includes several steps. First, approximately sixty demographic and attitudinal questions from the General Social Survey (GSS), a biennial face-to-face survey of a cross-section of US adults, are administered to a sample of respondents from each source under consideration (opt-in panels only, to our knowledge). Pinnacle then compares the resulting responses to those from the GSS, taking appropriate steps to ensure that the comparisons are fair (e.g., responses of eighteen to twenty-four year old men are compared with those of their counterparts who were interviewed in the GSS). Next, Pinnacle selects the combination of sources that balance, or cancel out, observed response differences between the sources under consideration and the GSS. For future surveys that use those sources, Pinnacle asks potential respondents a subset of questions from the GSS, then allocates them via a proprietary algorithm to the survey of interest. How they do so, according to Pinnacle, ensures comparability with the GSS and some degree of representativeness relative to the US general population as well.

Marketing Inc.’s Grand Mean Project is a different approach for selecting complementary (or interchangeable) sample sources. “Consistency is king” seems to be the premise underlying the approach. As Steve Gittleman of Marketing Inc. explained, “If I could monitor only one thing from here on out, it would be how consistent a panel is, the replicability of the data...Overall consistency of a panel’s output is the most important consideration in selecting an online sample provider” (“Panel Consistency…”, 2009, p. 1).

The Grand Mean Project administers a large battery of questions to respondents representing (ideally) all major sample sources in all countries in which researchers conduct online research. It then evaluates how mean responses from any single source, and segments within it, compare with those from other sources in order to identify complementary or interchangeable sources (and segments). According to Gittleman, the process ensures “that a shift in tracking study results is real and not caused by changes in the sample” (“Panel Consistency…,” 2009, p. 3). Gittleman identifies additional benefits as well.

To summarize, proponents of new sample source selection approaches cite at least three benefits: (a) consistency (or interchangeability) of new respondent sources with existing ones, (b) complementariness of new respondent sources with existing ones relative to an external standard, and (c) enhanced representativeness relative to the US general population (in Pinnacle’s case) through calibration with the GSS.
It is difficult to assess the merit of some of these claims, partly because peer-reviewed journals, such as *Public Opinion Quarterly* and the *International Journal of Market Research* contain no publications that describe or evaluate the underlying methodologies. This is not altogether surprising. Market researchers often conduct research as quickly as possible to attempt to meet important commercial needs. Unlike some of their colleagues working in academia, they do not usually feel the pressure to publish in peer-reviewed journals to advance their careers. For now, therefore, we consider these approaches to be important advances.

Our concern lies more with what they do not accomplish. Although they have taken a step in the right direction, we believe they have not gone far enough for at least two reasons: (a) they limit the supply of potential respondents, and (b) they rely on benchmark data sets that have either limited shelf lives or uncertain external validity. We therefore suspect that they may reduce sample representativeness and response accuracy in comparison to a new methodology, which we refer to later as *Propensity Score Select*, that selects potential respondents for surveys, no matter their originating source, based on how well their characteristics match an appropriate, evolving standard with demonstrated evidence of external validity.

We recognize, of course, that *Pinnacle*, the *Grand Mean Project*, and *Propensity Score Select* may be of little interest to critics of online research who do not believe it is possible to draw inferences from surveys of respondents selected by means other than probability sampling. Although the critics’ names have changed through the years, the criticisms have largely remained the same. In 1999, for instance, Mitofsky maintained that no matter how researchers adjust the results of a survey among respondents selected by means other than probability sampling, they would not be able to correct for the biases that arise from the difference between the sample and the population of interest (specifically, the general population). As he asserted, “...the willingness to discard the use of sampling frames as a means of selecting a sample and then the feeble attempts at manipulating the resulting bias…undermine the credibility of the survey process” (Mitofsky, 1999, p. 26).

More recently, Krosnick contended that there is no theoretical justification for why an opt-in panel of potential survey respondents (selected by means other than probability sampling) can constitute a credible sampling frame for surveys that purport to represent the attitudes, opinions, and behaviors of a broader population. According to Krosnick, “...to draw a scientific and representative sample of all residents of city and regional areas, it would be necessary to use a procedure that gives every member of that population an equal (or known) probability of being selected to participate in the survey” (Krosnick, 2008, p. 8).

Mitofsky and Krosnick’s views represent conventional wisdom in sampling theory--their comments could have come directly from a textbook. Indeed, Leslie Kish, author of more than
one textbook, defines a probability sample as one that is selected through a procedure in which “...all elements in the population frame receive known positive probabilities of selection, which are operationally defined and not necessarily equal” (Kish, 1995, p. 11).

A few years ago, perhaps for the reasons stated above, the American Association for Public Opinion Research (AAPOR) began on its web site to refer to sample selection procedures that do not meet the standard Kish put forward as SLOP, an acronym for “self-selected opinion polls”. More recently, AAPOR softened its stance. At its 2011 annual conference, for instance, (then) AAPOR president Frank Newport encouraged members to adopt a more open posture towards certain new methods of estimating public opinion, such as “non-probability research” (Newport, 2011, p. 602). He did so after first drawing members’ attention to a paper presented that same day which suggested that “...a carefully executed Internet opt-in panel produces estimates that are as accurate as a telephone survey and the two modes differ little in their estimates of other political indicators and correlates” (Ansolabehere, Fraga, and Schaffner, 2011, p. 1).

Newport’s call for increased open-mindedness is both timely and practical for various reasons. Some evidence suggests (e.g., see Christian et al., 2010), for example, that a confluence of influences, including the shrinking landline sampling frame, declining response rates, and the higher costs associated with accessing and interviewing cell phone users are creating significant problems for public opinion researchers; notably, those who depend on random-digit-dialing and related telephony approaches to conduct interviews. Soon, the livelihood of some of those researchers may hinge on identifying or inventing a suitable alternative to probability sampling. There may be even more at stake, particularly if you believe that the most important function of public opinion polling is to let the voice of the public be heard.

This paper’s aim, given this landscape, is to explain how and why the *Propensity Score Select* methodology may be able to reduce or minimize bias between the selected and achieved sample, and many target populations. The latter might include the general population, the online population, past participants in copy testing or brand tracking programs, or even individuals selected at random to take part in survey research by telephone or face-to-face. The *Propensity Score Select* methodology is versatile and of possible interest both in the US and abroad to diverse groups. These groups include market, government, and public opinion researchers, and various types of researchers working in academia, such as statisticians, sociologists, psychologists, and political scientists.

In what follows, we begin by describing commonly used respondent selection methods. Then, after drawing attention to historical antecedents for developing variations to these methods, we introduce and describe the *Propensity Score Select* methodology. Next, we present
empirical evidence bearing on its effectiveness. Finally, we consider possible implications of Propensity Score Select on methods other than online research with non-probability samples.

2. Random Sampling

There are many different methods available today for selecting respondents for online surveys. The simplest method may be to select a random sample from the population to which you have access (i.e., “the accessible population”) through panels, rivers, and other common sources. If the accessible population has any demographic, attitudinal, or behavioral skews in comparison to the target population, however, then a random sample would reproduce those skews, all other factors being equal. For instance, if women comprise 52 percent of the target population (e.g., all US adults, eighteen and older) but 26 percent of the accessible one, then women would also comprise 26 percent of a random sample, or half the number that researchers typically desire or require to represent the target population.

Even when the accessible and target populations are one and the same, a random sample of the accessible population may look different from the sub-sample that eventually completes the particular survey (i.e., “the achieved sample”) because of non-response (e.g., non-contact, unwillingness to participate) and non-random attrition within the survey.

In practice, few researchers are interested in drawing inferences to the online population to which they have access because it does not include all individuals, or all individuals who use the Internet. Nor is it a random sample of all Internet users. Instead, it is a list (more or less) of non-randomly selected people of people who have decided to participate in online surveys in response to offers from organizations that develop such lists.\(^1\) For these reasons, it seems farfetched to think that is possible to draw inferences about the entire population (including non-online users) through surveys of people on that list. Put even less delicately, the mere notion offends the sensibilities of many reasonable people. That may be why some organizations, such as the Associated Press and ABC News, refuse to publish or promote the results of such surveys, no matter how compelling their content. Not everyone is in agreement, however. For instance, media organizations, such as the Financial Times, are less restrictive than the Associated Press and ABC News and publish the results of such surveys routinely. Many global corporations, including the Walt Disney

\(^1\) A point to note here is that the concepts of sampling error and non-response error are not particularly meaningful to researchers who select respondents for surveys from incomplete lists that represent nothing but themselves (at least in a sampling science sense). The response rate is important to them, however, because it is a direct indicator of respondent supply.
Company, which owns ABC News, fall into the less restrictive camp as well. They have depended on the results of online survey research to inform strategic and tactical decision making for many years. Even in the same organization, attitudes towards online research can differ.

2.1. Standard Quota Sampling

Most researchers who believe it is possible to produce credible information through online research rely on some form of quota sampling, rather than random sampling, to select respondents for surveys. They often begin by dividing the target population into a set of mutually-exclusive groups (e.g., men, women, eighteen to twenty-four years old, twenty-five to twenty-nine years old) before they then specify how many respondents to recruit from each group. In general, once particular quota groups have been filled, potential respondents who would have otherwise qualified for the survey would be turned away. Alternatively, a routing system might direct them to another survey.

The specific manner in which researchers implement quota sampling can vary. For instance, they might set quotas on individual demographic variables (e.g., age or gender), on levels within those variables (e.g., on age groupings), on the interaction between those variables (e.g., age and gender) and the levels within them (e.g., men, women, eighteen to twenty-four years old), on survey stages (e.g., invitations, starts completes), on sample sources (e.g., the Toluna panel, the Research Now panel), on individual respondent attitudes or beliefs (e.g., attitudes towards technology, political affiliation), or on the many possible combinations among all these factors.

They might also take shortcuts and depend on judgment to fill difficult-to-fill quotas faster. For instance, they might offer higher incentives to less-frequent responders, such as eighteen to twenty-four-year-old men. Or, they might elect not to verify through third party databases (a standard procedure in online research today) the identities of members of hard-to-reach groups, such as African-Americans or Hispanics attending universities.

Despite the different ways in which researchers implement quota sampling, it is clear that quotas can help to make achieved samples look more like target populations. It is also clear that looks can be deceiving, particularly when the quotas do not encompass all relevant characteristics of those populations. Quotas set on individual characteristics such as age, gender, and income rather than on the interaction among the three (and the levels within them), for example, may produce skews (e.g., too many non-working people) that affect sample representativeness and response accuracy. More generally, it is difficult to capture the joint
distribution of all relevant respondent characteristics when using the types of quotas described thus far.

2.2. Quotas as a Set of Probabilities

A different way to approach the matter is to conceive of quotas as a set of probabilities representing the predilection of each potential survey respondent, given his or her demographic, attitudinal, and behavioral characteristics, to be a member of the target population rather than an alternative one. As described elsewhere (Terhanian et al., 2001; Terhanian, 2008), parallel surveys in combination with logistic regression can be used to produce each respondent’s predilection, or probability. Statistical theory, in turn, would help to ensure that individuals who have the same probability also have the same joint distribution on all variables in the regression model.

3. Historical Antecedents of the “Parallel Surveys” Approach

The thinking behind the “parallel surveys” approach can be traced back nearly sixty years to the work of statisticians William Cochran, John Tukey, and Frederic Mosteller (1954). After acknowledging that at times it is difficult to interview large numbers of respondents through probability sampling, they proposed parallel surveys to measure the sexual behavior of adult males, in their review of the methodology of the Kinsey Report (Kinsey, Pomeroy and Martin, 1948).

“Since it would not have been feasible for KPM to take a large sample on a probability basis,” they wrote, “a reasonable probability sample would be, and would have been, a small one and its purpose would be: (1) to act as a check on the large sample; and (2) possibly to serve as a basis for adjusting the results of the large sample” (Cochran, Tukey, and Mosteller, 1954, p. 23).

Cochran, Tukey, and Mosteller did not offer more detailed advice on how Kinsey and colleagues might have combined, or drawn inferences, from the parallel surveys, possibly because trustworthy tools for those purposes had not yet been invented. Nevertheless, they did offer hope to some researchers who work with non-probability samples.

As the years passed, demand for such tools did not wane. More than three decades later, for instance, a group of elite statisticians discussed (Wainer, 1986), among other topics, how it might be possible to draw inferences on the performance of the US educational system through student scores on the Scholastic Aptitude Test (SAT). They did so partly in response to assertions from officials in the Reagan Administration that reductions in spend-per-pupil had
contributed to higher SAT scores. Apparently, the officials did not mention that those same policies may have reduced the desire of less affluent and poor-performing students to attend college or to take the SAT, with those decisions perhaps having led to higher mean SAT scores. During discussion, the statistician John Tukey, who had become famous in some circles by that time, remarked, “I came in knowing little of this talk, and thus I was reminded of the story of there being two kinds of lawyers; there are the lawyers who tell you that you can’t do it and there are the lawyers who tell you how you can do it” (Wainer, 1986, p. 24).

After intimating that there were also corresponding categories of statisticians, Tukey demonstrated his own proclivity and suggested using the scores of statewide tests, which were administered to nearly all students, as a possible way to calibrate the scores from the SATs, which were administered to a self-selected sample of students. Tukey’s suggestion was in line with the one made to Kinsey and colleagues three decades earlier. As in 1954, Tukey stopped short of offering more detailed advice on how to build a bridge between the two instruments.

At about the same time, Paul Rosenbaum and Don Rubin (1983, 1984) invented a selection bias modeling technique known as propensity score adjustment to enable researchers to understand and possibly adjust for the differing reasons why people choose to engage in certain behaviors, activities or programs (such as the SAT). To understand the impact of cigarette smoking on the occurrence of lung cancer or on length-of-life, for instance, it is not possible for ethical or practical reasons to assign babies randomly at birth to a lifetime of smoking or non-smoking and then to compare them over time. It is possible, however, to compare later in life or even at death self-selecting groups of smokers and non-smokers with identical, or nearly identical, characteristics (if adequate data are available).

3.1. Making a Fair Comparison

According to Rosenbaum and Rubin, the propensity score, which one can estimate through logistic regression, reflects the predilection (in the form of a probability between 0 and 100 percent) of individuals to belong to one group rather than another given their characteristics. It can also serve as a means of matching one individual--a smoker--with another--a non-smoker--on all other characteristics.

To make a fair comparison between many smokers and non-smokers, the researcher might first sort smokers into quintiles based on their propensity score. The researcher might then use the same cut-off points to sort non-smokers into quintiles. In theory, smokers and non-smokers within the same quintile would then share the same joint distribution of all other observed characteristics, and the probability of being in that quintile would be equivalent for both groups. If it turns out that smokers in a particular quintile lived, say, ten years less than their
non-smoking counterparts in the same quintile, then one possible interpretation might be that smoking reduces life expectancy by ten years, on average, for like groups.

The researcher might then run the same analyses for the four remaining quintiles before computing an overall effect of smoking by taking the weighted average of the five. The evidence would be superior to anecdote or assertion but limited by the character and quality of the information in the statistical model. Nevertheless, in the words of Donald Campbell, “where randomized treatments are not possible...we must do the best we can with what is available to us” (Campbell, 1969, p. 411).

Campbell, and others with whom he collaborated, is perhaps most responsible for advancing the notion that it is possible to make fair comparisons among individuals (and groups) with identical, or nearly identical, characteristics as a precursor to possibly estimating causal effects (e.g., see Campbell and Stanley, 1963; Cook and Campbell, 1979). Campbell and colleagues were also vigilant (e.g., see Campbell and Boruch, 1975) to point out the limitations of methods other than random assignment (or “randomized experiments”) for that purpose. They warned that such methods do not control for unobserved differences between matched individuals or groups.

4. Inverting the Analytical Focus of Matching Methodologies

The idea of inverting the analytical focus of matching methodologies, such as propensity score adjustment, to improve survey design may lie with Boruch and Terhanian. They made the suggestion in 1996 in work commissioned by the US Department of Education (Boruch and Terhanian; 1996, 1998). At the time, Boruch and Terhanian observed that education surveys sponsored by the US federal government rarely explored individual’s intentions or reasons for membership in a program or group. They also noted that at times researchers would use data from those surveys (in addition to, or instead of, randomized controlled experiments and other techniques) to attempt to estimate the effects of program or group membership, such as the effects of dropping out of high school, on various outcomes. Those efforts were not as successful as they could have been, however, because the government had not designed its education surveys to enable researchers to estimate program or group effects.

In light of then-recent analytical advances such as propensity scoring, Boruch and Terhanian expected interest in conducting such analyses to increase. They therefore advised the National Center for Education Statistics (NCES) to consider transforming certain surveys into vehicles for estimating the effects of particular programs or groups, thereby expanding both the mission and possible value of those surveys. If NCES were to move forward, then some of those surveys would have required a significant re-design. Of course, NCES would have first needed to gather
from numerous stakeholders the necessary approvals, which would have been hard-to-come-by for various political and bureaucratic reasons.

In practice, applying tools such as propensity scoring to improve survey design can be challenging, not only for political or bureaucratic reasons but also because researchers have typically used such tools opportunistically to re-analyze existing data sets. That would probably not surprise Leslie Kish, who observed that “...over 95 percent of statistical attention in academia, textbooks, and publications is devoted to mathematical statistical analysis and only 2 percent to design” (Kish, 1995, p. 14). As one consequence, however, there are few examples or case studies on which to draw for those interested in using propensity scoring to improve survey design. We describe one example, which we refer to as “The Harris Interactive Case Study”, in the next section.

4.1. The Harris Interactive Case Study

To illustrate how one might invert the analytical focus of tools such as propensity scoring to improve survey design; notably, to make it easier to draw inferences from non-probability samples, consider the approach that we developed for Harris Interactive in 1999 (see Terhanian, 2008; Terhanian et al., 2001). In the intervening years, Harris has used the approach to try to contend with the many obstacles (such as non-coverage, self-selection, and the effects of panel membership on survey responses) that stand in the way of making population inferences through online research. The approach is akin to the one Cochran, Tukey, and Mosteller described in 1954. In short, Harris uses information from probability surveys of the general population as a check and as a means of adjusting data it collects through surveys of members of its opt-in panel.

4.2. The Harris Interactive Methodology

Harris conducts parallel (i.e., same questions asked at the same time) telephone or face-to-face and online surveys periodically. It views the exercise as a non-randomized experiment with a treatment (i.e., the online survey) and control group (i.e., the telephone or face-to-face survey) and multiple dependent variables; specifically, the responses to many of the survey questions (e.g., “If the election were held tomorrow, for which one of the following candidates [or parties] would you vote?”). Harris also includes within those surveys other questions (i.e., independent variables) that, it believes, will reduce or eliminate key differences between the two groups (and, more generally, between the online sample and the target population).

After Harris completes data collection, it weights the telephone or face-to-face data and the online data to national census targets, when applicable, before combining the two data sets.
Next, it estimates each respondent’s propensity score (i.e., the probability of having participated in the telephone or face-to-face survey rather than the online survey, given the variables in the model) through logistic regression, before comparing the propensity score distributions of the two samples. Harris might then use the propensity score, as well as other factors, in a second weighting procedure to bring proportions of key variables in the online sample in line with those of the telephone or face-to-face sample (and the target population).

For future online surveys when its interest lies in drawing inferences to, say, the general population, Harris would include the questions required to estimate each respondent’s propensity score in each online survey. It would then use the resulting score, and other factors, to attempt to reduce bias through weighting. The general population propensity score distribution it developed previously would serve as one benchmark or weighting target.

Harris can run into trouble in those later surveys, however, when the propensity score distribution of the achieved sample does not overlap with that of the benchmark or target population sample. To avoid such a predicament, we believe it makes good sense to select respondents for surveys based on how well their individual characteristics (and their joint distribution), as represented by the propensity score, match those of the target population. To be clear, we are proposing to shift the estimation of the propensity score from the data analysis stage of online surveys to the respondent selection stage.

The decision to make the shift would benefit from the low cost of asking questions online and the ability to re-use information (e.g., gender) that opt-in panelists and possibly others provided previously. The increasing use of respondent sources other than opt-in panels and the use of routers would provide additional benefits. When researchers rely on rivers as a sample source, for instance, they will not know much about potential respondents in advance. By placing the questions they need to estimate the propensity score at the survey’s “front door”, they will reduce the risk of including the wrong people (and the risk of not including enough of the right people).

There are possible downsides to the shift as well. Our experience suggests, for instance, that researchers may need to screen roughly two to three times as many individuals at the survey’s “front door” as compared to a methodology using quotas based on some combination of region, age, and gender. We attribute the difference to the difficulty of identifying respondents who are least likely to be a member of the “treatment” group (i.e., those with the lowest propensity score). A reader may wonder whether this would place a strain on the available supply of potential respondents. All else equal, it would do so. But all else is not equal. The new methodology focuses on the characteristics of individuals, not those of respondent
sources, to decide who to select for particular surveys. It therefore opens the door to respondents who are directed to a survey through any source. This is one reason why the “front door” methodology may make obsolete the more restrictive, source-based methodologies described earlier.

5. The Propensity Score Select Methodology

We describe below the specific steps—including the development of the propensity score—that are involved in implementing a “front door” methodology. From here on, we refer to the methodology as Propensity Score Select.

Step 1: Prepare to conduct parallel surveys in which you ask the same questions at roughly the same time to representative samples of the accessible and target populations. It may be helpful to regard the exercise as a non-randomized experiment with a treatment (i.e., the sample of respondents from the “accessible population”) and a control group (i.e., the sample of respondents from the “target population”), several independent variables (which different researchers might refer to as covariates, balancing variables, or variables “on the right side of the equation”), and several outcome variables. For this illustration, we will regard the US general population as the target population.

As we noted in section 2.1, if the control group is a random sample of the target population, it may still not represent that population accurately because of non-response, mode effects, or other sources of error. The same principle applies to the treatment group. Researchers should therefore exercise care and caution when selecting samples for the parallel surveys.

A second point to note is that the control group need not be a random sample of a country’s general population. It could be a random sample of its online population or its population of, say, iPad2 owners. It could even be the entire group of individuals interviewed previously in a tracking study or concept testing program, among many other possibilities. The ability to choose different target populations, and, therefore, different calibration samples or control groups, distinguishes the Propensity Score Select methodology from those described earlier. As a reminder, the GMI Pinnacle system relies exclusively on the GSS for calibration while the Grand Mean Project relies on a fixed set of proprietary questions and emphasizes consistency rather than validity or accuracy as its main goal.

Step 2: Choose questions to represent the independent variables. Ideally, the questions will account for all observed and unobserved differences between the control and treatment groups. Possible questions include demographic measures such as age, gender, region, education level, and others measured on the Census or the Current Population Survey. Possible non-demographic questions include attitudes towards privacy, attitudes towards security and
risk, television viewing behavior, measures of physical activity, and selected attitudes, behaviors and opinions measured on surveys such as the GSS, the American Community Survey (ACS), the National Health Interview Survey (NHIS), and the National Health and Nutrition Examination Survey (NHANES). Any question that is selected should include response choices that are applicable both for control and treatment group members.

Despite a large increase in propensity scoring’s popularity since the mid 1990s (for an estimate see Arcenaux, Gerber, and Green, 2010), it has been used chiefly as an analysis tool rather than a survey design tool, as we noted earlier. As a result, it is difficult to draw on a body of literature or evidence for advice on which variables to include in the model. For this reason, some initial research—perhaps even a great deal—may be needed to understand whether, how and why treatment group members differ from their counterparts in the control group. As we suggested in Section 4, however, it can be challenging to figure out how individuals make their way into one group rather than another. They can take many different pathways. We therefore agree, at least broadly, that “...trying to do a reasonable specification for a full set of pathways that might be important and correlated to a variety of outcome measures is exceedingly difficult” (Brick, 2011, p. 883). However, we see this as a starting point rather than a dead end. If we were lawyers, we suspect we would be of the second type that Tukey described: “the lawyers who tell you how you can do it”.

**Step 3:** Next, choose questions to represent outcome measures, which can be used to validate, and possibly calibrate, the model that is developed later. Ideally, the measures will be verifiable; that is, you will know, or can learn, whether they are accurate or true without the help of an omniscient observer. Possibilities include measures representing sales forecasts, past voting behavior (or, depending on timing, candidate or party preference for an imminent election), possession of a driver’s license, possession of a passport, and certain buying behaviors. Attitudes, behaviors and opinions measured on surveys such as the GSS, ACS, NHIS, NHANES, and possibly others, including certain telephone and face-to-face omnibus surveys, are options here as well.

It is important to choose questions that contain as little error as possible. Excluding questions with responses that are potentially embarrassing or socially desirable is in line with the recommendation, particularly if those questions were administered aurally or in the presence of an interviewer. It may be necessary to make some trade-offs. In order to include current information in the modeling, for instance, it may be necessary to rely on a non-government funded study, such as a telephone or face-to-face omnibus survey conducted by a market research company.

**Step 4:** After fielding the parallel surveys, use logistic regression to estimate each respondent’s probability of being a member of the control group rather than the treatment group. The
demographic and non-demographic questions selected in the second step would serve as the model’s independent variables. The best model would minimize the differences on the outcome measures (selected in the third step) among control and treatment group members with the same, or similar, propensity scores. In the interest of economy, it may be prudent to drop certain variables from the model.

A key output of the exercise would be a distribution of propensity scores for the control group, whatever it may be, which researchers can then use to decide which respondents to select for future surveys.

**Step 5:** Exercise good judgment when deciding, in particular, whether it is possible to represent the attitudes, experiences, and behaviors of the general population—both online and non-online users—through surveys of respondents selected through *Propensity Score Select*. Even if you happen to agree that “‘he did not use a probability sample’ is...not a criticism which should end further discussion and doom the inquiry to the cellar” (Cochran, Mosteller, and Tukey, 1954, p. 328), it is still sensible to acknowledge that online research among non-probability samples may not be suitable at times. For instance, if the aim is to estimate the percentage of homeless adults in Detroit, or to explore other matters correlated with online usage, then methods other than online research are a better option.

**Step 6:** As part of the respondent selection process for future surveys, first ask potential respondents the questions needed to estimate the propensity score (or, “pass through” to the survey previously collected and stored information from opt-in panelists to minimize burden on them). Next, assign each respondent, irrespective of his or her originating sample source, to available surveys based on the propensity score (which can be estimated in real time) until all propensity score quotas have been filled. In theory, the joint distribution of the characteristics of respondents directed to those surveys should match those of the target population, even when it includes non-online users. Moreover, the responses those individuals give to new survey questions should compare favorably to those that a representative sample of the target population would have given had it been selected through probability sampling.

**Step 7:** The *Propensity Score Select* methodology can and, perhaps, should be used with other techniques or methodologies, such as weighting (e.g., raking, cell weighting, propensity score weighting), to create possible efficiencies, to reduce bias further, or both. For instance, if there are imbalances in the propensity score distribution for any number of reasons (e.g., differential drop outs by quintile or decile, implementation problems, time constraints), then weighting might help to reduce bias. Understanding how to develop or identify the optimal weighting solution when using propensity scores is both a theoretical and empirical question that has received some constructive attention recently from academic researchers (e.g., see Lee and Valliant, 2009; Valliant and Dever, 2011). Our own experience suggests that weighting on the
propensity score as well as on various socio-demographic characteristics will reduce more bias than weighting on the propensity score alone.

**Step 8 (and beyond):** The process described thus far is an ongoing one. Any model that is developed will not stand the test of time. The attitudes, experiences, and behaviors of target populations will change so the model development process described here will need to be repeated periodically to account for such change. Implementation can also present challenges. For instance, Harris Interactive produced a final forecast for the 2004 US presidential election that was based on an incorrect implementation of its (related) propensity score weighting methodology. The company did not calibrate (e.g., see Step 3 above) its model correctly, as it had in 2000, which caused it to produce an off-target forecast. Harris corrected the problem in time for the 2005 British General Election, for which it produced an accurate forecast. Later Harris forecasts in 2008 (US Presidential election) and 2010 (British General election) were on target too.

6. **New Empirical Evidence**

In what follows, we will evaluate the effectiveness of the Propensity Score Select methodology using data from a fifteen-minute survey that explored various topics, including general issues, attitudes towards privacy, technology ownership, and online behaviors. We describe the topics below in more detail. Hereafter, we refer to them as “content” questions.

1. **General Issues:** Quality of health, Approval for President Obama, Degree of religiousness, Own passport, Possess driver’s license, Smoke cigarettes, TV viewing per week

2. **Attitudes Towards Privacy:** AIDS screening at work, Unsolicited calls for selling purposes, Cookies on computers for tracking purposes, Airport searches based on visual profiles

3. **Technology Ownership:** Smartphone, Digital Camera, Tablet Computer, Game Console, Satellite Radio, eBook Reader

4. **Online Behaviors** (since 1/1/11): Made purchase, Banked, Used social network/media Application, Uploaded picture, Watched video, Participated in auction

The questionnaire was administered in late February and early March, 2011 via two modes (telephone and online) to four samples (one telephone, three online). For the telephone sample, quotas were established for region and gender. For the online samples, quotas were established not only for region and gender but also for age, race-ethnicity, and education level. Table 1 shows the quota targets for the online samples.
Table 1. Quota Targets for the Online Samples

<table>
<thead>
<tr>
<th></th>
<th>Total (1100)</th>
<th>Male (550)</th>
<th>Female (550)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northeast</td>
<td>19%</td>
<td>18%</td>
<td>18%</td>
</tr>
<tr>
<td>Midwest</td>
<td>22%</td>
<td>22%</td>
<td>22%</td>
</tr>
<tr>
<td>South</td>
<td>37%</td>
<td>37%</td>
<td>37%</td>
</tr>
<tr>
<td>West</td>
<td>23%</td>
<td>23%</td>
<td>22%</td>
</tr>
<tr>
<td>Ages 18-24</td>
<td>&gt;= 9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ages 25-34</td>
<td>&gt;= 12%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ages 35-44</td>
<td>&gt;= 12%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ages 45-54</td>
<td>&gt;= 13%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ages 55-64</td>
<td>&gt;= 10%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ages 65+</td>
<td>&gt;= 11%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>African-American</td>
<td>&gt;= 10%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>&gt;= 10%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than HS grad</td>
<td>&gt;= 11%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS grad</td>
<td>&gt;= 27%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some college</td>
<td>&gt;= 23%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>College grad</td>
<td>&gt;= 26%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We provide additional details on the four samples below:

- **Telephone (Landline):** 1019 completed interviews among respondents residing in private households in the continental US. Respondents were contacted through random-digit-dialing by the Opinion Research Corporation during the period February 24-27, 2011. The data were weighted to Census or CPS estimates to bring proportions for region, gender, age, race-ethnicity, education, household income, and “lived in home without landline in past two years” in line with those of the population.

- **Direct Invitation (Online):** 1100 completed interviews during the period March 2-6, 2011 among members of Toluna’s online opt-in panel who were invited directly to the survey by email. The data to be presented here were not weighted.

- **River (Online):** 1100 completed interviews during the period March 2-6, 2011 among non-members of Toluna’s online opt-in panel who were directed to the survey after clicking through an advertisement or link on the Internet. The data to be presented here were not weighted.

- **Router (Online):** 1098 completed interviews during the period March 2-6, 2011 among members and non-members of Toluna’s online opt-in panel who were routed to the
survey, typically after not having qualified for one or more other surveys. The data to be presented here were not weighted.

To estimate each respondent’s propensity score, we used questions representing region, gender, age, race-ethnicity, level of education, household income and several attitudes or opinions covering broad areas and issues that differentiate online users who take part in online surveys from members of the general population.

Although it is possible to estimate more than one propensity score for each respondent (e.g., one to reduce or eliminate differences between respondents and the general population, and a second to reduce differences between respondents and the online population), we elected to estimate only the general population score. For questions about online behaviors, the score is less than optimal but more than serviceable based on our experience.

6.1. Simulating the Effects of the Propensity Score Select Methodology

To simulate the effects of the Propensity Score Select methodology, a precursor to estimating its accuracy, we selected from the combined pool of online respondents a stratified random sample of 1000 distributed equally across a lone stratum; namely, the general population propensity score quintile distribution. The Propensity Score Select sample, therefore, is a sub-sample of respondents who are also represented in the direct invitation, river, and router samples.

7. Assessing Accuracy

To assess accuracy, we follow recent precedent (e.g., see Yeager et al., 2011) and regard the modal response of each question as the one of interest. For each source, we then calculate the (absolute) difference between each question’s modal response and the benchmark. We refer to the overall average of those differences as the “mean absolute deviation”.

The second, more important key measure is the Mean Squared Error (MSE). We consider the MSE to be an excellent measure for comparing two or more sets of estimates because it takes into account the mean absolute deviation of the estimates (i.e., the first measure) as well as the variability of those estimates. There is longstanding precedent for using the MSE for this purpose, dating back (to our knowledge) to the evaluation of the US pre-election polls of 1948 (Mosteller et al., 1949).

For the demographic and household questions, information from the 2010 Census, the Current Population Survey, or other government-funded surveys will serve as the benchmarks. For all but one content question (see footnote in Table 5), we have chosen to regard the responses from the telephone survey as the benchmarks. We recognize that any of these measures could
be biased for various reasons. We are more confident (obviously) in the accuracy of Census and other government benchmarks than the telephone ones.

In what follows, we consider, in turn, the accuracy of the responses to the five demographic questions that served as quotas, the four demographic and household questions that did not serve as quotas, and the twenty-three content questions. The sample (or selection method) with the lowest (average) score on the MSE will be considered the most accurate.

### 7.1. “Demographic, Quota” Questions

The information in Table 2 suggests that the quotas for region, gender, age, race-ethnicity, and education level were implemented correctly—the minimum targets that Table 1 reported were either achieved or nearly achieved. Table 2 suggests, as well, that individuals who were selected through the Propensity Score Select methodology gave responses that were the most accurate and the least variable. For these reasons, the Propensity Score Select MSE was 80% lower than that of Direct Invitation respondents, its closest competitor, and, at the other extreme, 632% lower than that of River respondents.

One measure that stands out in Table 2 is the comparatively high percentage of River respondents (37.9% versus a benchmark of 30.4%) who selected “high school” as the last grade in school they completed, or degree they received. This may be a result of the decision not to place caps on the education level quotas.

**Table 2. Accuracy of the Responses to “Demographic, Quota” Questions**

<table>
<thead>
<tr>
<th>Modal Response</th>
<th>Benchmark</th>
<th>Direct Invitation</th>
<th>River</th>
<th>Router</th>
<th>Propensity Score Select</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>Region, &quot;South&quot;</td>
<td>37.2</td>
<td>36.6</td>
<td>36.6</td>
<td>35.9</td>
<td>35</td>
</tr>
<tr>
<td>Gender, &quot;Female&quot;</td>
<td>51.4</td>
<td>50</td>
<td>50</td>
<td>52.4</td>
<td>53.1</td>
</tr>
<tr>
<td>Age Grouping, &quot;45-54&quot;</td>
<td>19</td>
<td>19.4</td>
<td>19.8</td>
<td>19.4</td>
<td>19.4</td>
</tr>
<tr>
<td>Race-Ethnicity, &quot;White, NH&quot;</td>
<td>67.8</td>
<td>66.5</td>
<td>71.4</td>
<td>72.7</td>
<td>68.8</td>
</tr>
<tr>
<td>Education, &quot;HS Degree&quot;</td>
<td>30.4</td>
<td>26.7</td>
<td>37.9</td>
<td>28.8</td>
<td>31.7</td>
</tr>
</tbody>
</table>

**Key Measures**

<table>
<thead>
<tr>
<th>Base</th>
<th>1100</th>
<th>1100</th>
<th>1098</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Absolute Deviation</td>
<td>1.5</td>
<td>2.8</td>
<td>1.8</td>
<td>1.3</td>
</tr>
<tr>
<td>Variance</td>
<td>1.7</td>
<td>8.4</td>
<td>3.1</td>
<td>0.5</td>
</tr>
<tr>
<td>MSE</td>
<td>4</td>
<td>16.1</td>
<td>6.5</td>
<td>2.2</td>
</tr>
<tr>
<td>MSE vs. PSS</td>
<td>-80%</td>
<td>-632%</td>
<td>-194%</td>
<td></td>
</tr>
</tbody>
</table>
7.2. “Demographic (Non-Quota) and Household” Questions

The information in Table 3 indicates that the Propensity Score Select MSE was 15% higher than that of River respondents and roughly 186% lower than that of both Direct Invitation and Router respondents. As a reminder, the four benchmark responses that are reported in Table 3 did not play a role (i.e., through quotas) in the original respondent selection process.

One measure that stands out in Table 3 is the comparatively high percentage of “married” respondents (59.3% versus a benchmark of 53.6%) in the Router sample. An obvious implication is that Router respondents are more likely to be married than others. Implications can have their own implications (e.g., it might be prudent to establish quotas on marital status when selecting respondents through the Router, or it might make sense to include marital status in a propensity score model) but we will not pursue this line of thinking further here.

Table 3. Accuracy of the Responses to “Demographic (Non Quota) and Household” Questions

<table>
<thead>
<tr>
<th>Modal Response</th>
<th>Benchmark</th>
<th>Direct Invitation</th>
<th>River</th>
<th>Router</th>
<th>Propensity Score Select</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>HH Income, &quot;$50-$74.99k&quot;</td>
<td>17.7</td>
<td>22.7</td>
<td>18.4</td>
<td>21</td>
<td>19.7</td>
</tr>
<tr>
<td>Marital Status, &quot;Married&quot;</td>
<td>53.6</td>
<td>58.2</td>
<td>52.3</td>
<td>59.3</td>
<td>53.3</td>
</tr>
<tr>
<td>Own or Rent, &quot;Own&quot;</td>
<td>66.2</td>
<td>68.6</td>
<td>61.9</td>
<td>68.9</td>
<td>64.2</td>
</tr>
<tr>
<td>Household Members, &quot;2&quot;</td>
<td>33.4</td>
<td>38.6</td>
<td>33.9</td>
<td>38.7</td>
<td>37.6</td>
</tr>
</tbody>
</table>

Key Measures

<table>
<thead>
<tr>
<th></th>
<th>Base</th>
<th>Direct Invitation</th>
<th>River</th>
<th>Router</th>
<th>Propensity Score Select</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>1100</td>
<td>1100</td>
<td>1098</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>Mean Absolute Deviation</td>
<td>4.3</td>
<td>1.7</td>
<td>4.3</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>1.7</td>
<td>3.1</td>
<td>2.2</td>
<td>2.6</td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>20.2</td>
<td>6</td>
<td>20.2</td>
<td>7.1</td>
<td></td>
</tr>
<tr>
<td>MSE vs. PSS</td>
<td>-185%</td>
<td>15%</td>
<td>-186%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7.3. Overall Accuracy of All “Demographic and Household” Questions

Table 4 summarizes the information reported in Tables 2 and 3. The summary suggests that the individuals who were selected through the Propensity Score Select methodology gave the most accurate responses across the nine measures—the Propensity Score Select MSE of 4.2 was 162% lower than the River MSE of 11. We recognize that buyers of market research, among others, are far more concerned with the accuracy of responses to content questions, a topic to which we now turn.
Table 4. Accuracy of the Responses to All “Demographic and Household” Questions

<table>
<thead>
<tr>
<th>All Demographic and Household Questions</th>
<th>Direct Invitation</th>
<th>River</th>
<th>Router</th>
<th>Propensity Score Select</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Key Measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>1100</td>
<td>1100</td>
<td>1098</td>
<td>1000</td>
</tr>
<tr>
<td>Mean Absolute Deviation</td>
<td>2.7</td>
<td>2.3</td>
<td>2.9</td>
<td>1.7</td>
</tr>
<tr>
<td>Variance</td>
<td>3.7</td>
<td>5.7</td>
<td>4</td>
<td>1.4</td>
</tr>
<tr>
<td>MSE</td>
<td>11.2</td>
<td>11</td>
<td>12.4</td>
<td>4.2</td>
</tr>
<tr>
<td>MSE vs. PSS</td>
<td>-168%</td>
<td>-162%</td>
<td>-198%</td>
<td></td>
</tr>
</tbody>
</table>

7.4. “General” Questions

Seven questions were included in the “General Questions” section. As Table 5 shows, responses from individuals selected through the Propensity Score Select methodology differed by 3.4 percentage points, on average, from the benchmark. The Propensity Score Select responses were also less variable than those from the other sources. For these reasons, the Propensity Score Select MSE was 49 percent lower than the Direct Invitation MSE, its closest competitor.

Table 5. Accuracy of the Responses to “General” Questions

<table>
<thead>
<tr>
<th>Modal Response</th>
<th>Benchmark</th>
<th>Direct Invitation</th>
<th>River</th>
<th>Router</th>
<th>Propensity Score Select</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>Health, &quot;Good&quot;</td>
<td>32.3</td>
<td>32</td>
<td>33.6</td>
<td>32.2</td>
<td>31.5</td>
</tr>
<tr>
<td>Obama, &quot;Approve&quot;</td>
<td>46.6</td>
<td>53.3</td>
<td>48.2</td>
<td>48.5</td>
<td>53.9</td>
</tr>
<tr>
<td>Religious, &quot;Moderately&quot;</td>
<td>40.2</td>
<td>37.2</td>
<td>39.2</td>
<td>36.1</td>
<td>38.1</td>
</tr>
<tr>
<td>Passport, &quot;Do not Own&quot;</td>
<td>56.7</td>
<td>55.6</td>
<td>61.5</td>
<td>56</td>
<td>55.8</td>
</tr>
<tr>
<td>Driver's License, &quot;Have&quot;</td>
<td>85.5</td>
<td>88.4</td>
<td>84.8</td>
<td>88.1</td>
<td>86.4</td>
</tr>
<tr>
<td>Smoke Cigarettes, &quot;Not at all&quot;</td>
<td>80.7</td>
<td>73.5</td>
<td>64</td>
<td>68.6</td>
<td>77.8</td>
</tr>
<tr>
<td>TV, &quot;2 hrs per Week&quot;</td>
<td>25.5</td>
<td>15.5</td>
<td>14.6</td>
<td>15.2</td>
<td>16.9</td>
</tr>
<tr>
<td><strong>Key Measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>1019</td>
<td>1100</td>
<td>1100</td>
<td>1098</td>
<td>1000</td>
</tr>
<tr>
<td>Mean Absolute Deviation</td>
<td>4.5</td>
<td>5.3</td>
<td>4.5</td>
<td>3.4</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>12.7</td>
<td>38.5</td>
<td>22.6</td>
<td>10.6</td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>32.6</td>
<td>66.4</td>
<td>43.3</td>
<td>21.8</td>
<td></td>
</tr>
<tr>
<td>MSE vs. PSS</td>
<td>-49%</td>
<td>-204%</td>
<td>-98%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: The National Household Interview Survey (2010) is the source of the percent of adults who report that they smoke cigarettes “not at all”.*
7.5. “Attitudes Towards Privacy” Questions

The “Attitudes Towards Privacy” section included four questions. Responses from individuals selected through the Propensity Score Select methodology differed by 2.4 percentage points, on average, from the benchmark, as Table 6 shows. As with the “General Questions”, the Propensity Score Select set of responses was less variable than the others. As a result, the Propensity Score Select MSE was 419 percent lower than its closest competitor.

Table 6. Accuracy of the Responses to “Attitudes Towards Privacy” Questions

<table>
<thead>
<tr>
<th>Modal Response</th>
<th>Benchmark</th>
<th>Direct Invitation</th>
<th>River</th>
<th>Router</th>
<th>Propensity Score Select</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>AIDS Screening, &quot;Violation&quot;</td>
<td>53.5</td>
<td>49.6</td>
<td>43.2</td>
<td>47.6</td>
<td>51.4</td>
</tr>
<tr>
<td>Unsolicited Calls, &quot;Violation&quot;</td>
<td>67.3</td>
<td>78.3</td>
<td>76.9</td>
<td>77.8</td>
<td>69.2</td>
</tr>
<tr>
<td>Cookies, &quot;Violation&quot;</td>
<td>65.8</td>
<td>71.3</td>
<td>72.1</td>
<td>70.2</td>
<td>70.7</td>
</tr>
<tr>
<td>Airport Search, &quot;No Violation&quot;</td>
<td>60.4</td>
<td>59.7</td>
<td>59.2</td>
<td>64.5</td>
<td>61.2</td>
</tr>
</tbody>
</table>

Key Measures

| Base                           | 1019      | 1100              | 1100  | 1098   | 1000                    |
| Mean Absolute Deviation        | 5.3       | 6.9               | 6.2   | 2.4    |                         |
| Variance                       | 18.5      | 17.2              | 8.7   | 3      |                         |
| MSE                            | 46.4      | 64.2              | 47.5  | 8.9    |                         |
| MSE vs. PSS                    | -419%     | -618%             | -432% |        |                         |

7.6. “Technology Ownership” Questions

The “Technology Ownership” section included six questions. As Table 7 shows, responses from individuals selected through the Propensity Score Select methodology differed by 2.3 percentage points, on average, from the benchmark. The Propensity Score Select responses were less variable than those from Direct Invitation and Router sources but slightly more variable than those from the River. The Propensity Score Select MSE was the lowest of all by a slim margin.
Table 7. Accuracy of the Responses to “Technology Ownership” Questions

<table>
<thead>
<tr>
<th>Modal Response</th>
<th>Benchmark</th>
<th>Direct Invitation</th>
<th>River</th>
<th>Router</th>
<th>Propensity Score Select</th>
</tr>
</thead>
<tbody>
<tr>
<td>Own Smartphone, &quot;No&quot;</td>
<td>58.8</td>
<td>61.6</td>
<td>64.7</td>
<td>60.8</td>
<td>63.2</td>
</tr>
<tr>
<td>Own Digital Camera, &quot;Yes&quot;</td>
<td>73.4</td>
<td>82.3</td>
<td>77.6</td>
<td>82.1</td>
<td>80.5</td>
</tr>
<tr>
<td>Own Tablet, &quot;No&quot;</td>
<td>91.4</td>
<td>93.1</td>
<td>92.2</td>
<td>91.6</td>
<td>92.2</td>
</tr>
<tr>
<td>Own Game Console, &quot;No&quot;</td>
<td>53</td>
<td>50.2</td>
<td>53.1</td>
<td>52.2</td>
<td>52.9</td>
</tr>
<tr>
<td>Own Satellite Radio, &quot;No&quot;</td>
<td>78.5</td>
<td>79.9</td>
<td>82.9</td>
<td>77.4</td>
<td>79.2</td>
</tr>
<tr>
<td>Own eBook Reader, &quot;No&quot;</td>
<td>90</td>
<td>88.4</td>
<td>88.7</td>
<td>87.8</td>
<td>90.7</td>
</tr>
</tbody>
</table>

Key Measures

<table>
<thead>
<tr>
<th></th>
<th>Base</th>
<th>1019</th>
<th>1100</th>
<th>1100</th>
<th>1098</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Absolute Deviation</td>
<td>3.2</td>
<td>2.8</td>
<td>2.5</td>
<td>2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>8.2</td>
<td>5.5</td>
<td>9.8</td>
<td>7.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>18.4</td>
<td>13.3</td>
<td>16</td>
<td>13.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSE vs. PSS</td>
<td>-39%</td>
<td>0%</td>
<td>-21%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The evidence presented in Table 7, and possible implications of this evidence (e.g., it may be safe to use an online approach to estimate the percent of adults who own different technology products), appears to contradict advice given by Duffy and colleagues (2005).

According to Duffy and colleagues, “...There are certain survey questions that will never be appropriate to ask online when you are trying to represent the population as a whole--and technology use is certainly one of those. The entire sample source by the very nature of the survey approach has Internet access of some sort, and internet access correlates very highly with many types of technology usage--particularly computer ownership” (Duffy et al., 2005).

Given that they carried out their research in Great Britain in 2003, it appears that their recommendation has not fully stood the tests of time and geography. This is not surprising given the increasing adoption of the Internet over the past several years, among other possible reasons.

7.7. “Online Behaviors” Questions

Six questions were included in the “Online Behaviors” section. As Table 8 shows, responses from individuals selected through the Propensity Score Select methodology differed by 5.1 percentage points from the benchmark, the lowest difference of all shown. The Propensity Score Select MSE was 7 percent lower than the MSE of River respondents, its closest competitor. As we noted earlier, we developed the propensity score used here to reduce or eliminate differences between online respondents and the general population, not its online
sub-population. Nevertheless, the evidence suggests that the Propensity Score Select methodology is comparatively effective at reducing bias in responses to questions about online behaviors.

Table 8. Accuracy of the Responses to “Online Behaviors” (since Jan 1, 2011)

<table>
<thead>
<tr>
<th>Modal Response</th>
<th>Benchmark</th>
<th>Direct Invitation</th>
<th>River</th>
<th>Router</th>
<th>Propensity Score Select</th>
</tr>
</thead>
<tbody>
<tr>
<td>Made Online Purchase</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>Banked Online</td>
<td>58.1</td>
<td>73.6</td>
<td>64.2</td>
<td>72.1</td>
<td>68.6</td>
</tr>
<tr>
<td>Used Social Media</td>
<td>64.1</td>
<td>71.3</td>
<td>72.1</td>
<td>66.8</td>
<td>71.4</td>
</tr>
<tr>
<td>Uploaded Picture</td>
<td>61.8</td>
<td>57.1</td>
<td>54.4</td>
<td>57.4</td>
<td>57.8</td>
</tr>
<tr>
<td>Watched Video Online</td>
<td>70.9</td>
<td>72.5</td>
<td>73.8</td>
<td>75.5</td>
<td>75.4</td>
</tr>
<tr>
<td>Participated in Online Auction</td>
<td>86.4</td>
<td>73.9</td>
<td>78.5</td>
<td>76.1</td>
<td>83.7</td>
</tr>
<tr>
<td><strong>Key Measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mean Absolute Deviation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>846</td>
<td>1100</td>
<td>1100</td>
<td>1098</td>
<td>1000</td>
</tr>
<tr>
<td><strong>Variance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MSE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MSE vs. PSS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7.8. Accuracy of All “Content” Questions

Overall, we evaluated twenty-three questions across the survey’s four main content sections. Table 9 shows that responses from individuals selected through the Propensity Score Select methodology differed by 3.4 percentage points, on average, from the benchmark. The table indicates, as well, that the Propensity Score Select MSE was 103 percent lower than the Router MSE, its closest competitor.

In contrast to earlier tables, Table 9 reports the percentage of respondents within each propensity score quintile. It shows, for instance, that approximately four in ten Direct Invitation, River, and Router respondents are members of Quintile 1 versus an expectation of two in ten. In addition, Quintile 2 is over-represented in those three sources, while Quintiles 3, 4 and 5 are under-represented. The Propensity Score Select methodology contributed to a substantial reduction in bias by selecting a lower percentage of individuals from Quintiles 1-2, and a higher percentage from Quintiles 3-5.
Table 9. Accuracy of the Responses to All "Content" Questions

<table>
<thead>
<tr>
<th>All Content Questions</th>
<th>Benchmark</th>
<th>Direct Invitation</th>
<th>River</th>
<th>Router</th>
<th>Propensity Score Select</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key Measures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>1019</td>
<td>1100</td>
<td>1100</td>
<td>1098</td>
<td>1000</td>
</tr>
<tr>
<td>% Quintile 1</td>
<td>20%</td>
<td>38.5%</td>
<td>40.1%</td>
<td>39.6%</td>
<td>20%</td>
</tr>
<tr>
<td>% Quintile 2</td>
<td>20%</td>
<td>26.3%</td>
<td>24.2%</td>
<td>25.5%</td>
<td>20%</td>
</tr>
<tr>
<td>% Quintile 3</td>
<td>20%</td>
<td>17.8%</td>
<td>18.6%</td>
<td>18.2%</td>
<td>20%</td>
</tr>
<tr>
<td>% Quintile 4</td>
<td>20%</td>
<td>10.1%</td>
<td>10.3%</td>
<td>10%</td>
<td>20%</td>
</tr>
<tr>
<td>% Quintile 5</td>
<td>20%</td>
<td>7.4%</td>
<td>6.8%</td>
<td>6.6%</td>
<td>20%</td>
</tr>
<tr>
<td>Mean Absolute Deviation</td>
<td>5.2</td>
<td>5</td>
<td>4.9</td>
<td>3.4</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>17.2</td>
<td>18.1</td>
<td>16.9</td>
<td>8.8</td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>43.7</td>
<td>42.9</td>
<td>40.9</td>
<td>20.2</td>
<td></td>
</tr>
<tr>
<td>MSE vs. PSS</td>
<td>-116%</td>
<td>-112%</td>
<td>-103%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8. Next Steps and Closing Thoughts

The evidence presented here suggests that the Propensity Score Select methodology enables market researchers to select online survey respondents who provide information that is more accurate and less variable than information provided by respondents selected by other means. Put somewhat differently, the evidence suggests that Propensity Score Select is more effective than a type of source selection (coupled with standard quota sampling) at ensuring sample representativeness and achieving response accuracy.

If this proves to be the case over time and across topics, study types, and geographies, then researchers clearly will need to spend more time working to develop and improve respondent selection procedures and less time tinkering with source selection. Market research buyers will need to change in some ways too. Today, some spend a great deal of time debating the relative merits of the three online sample sources considered here—Direct Invitation, River, and Router—no matter how respondents are selected from them. At the very least, those buyers will need to begin asking market research sellers new questions. Time spent choosing the best respondent source or the optimal combination (e.g., “Should we use the following three respondent sources with the first contributing half of all interviews and the second and third contributing a quarter each?”) may be time better spent, if not time wasted entirely. Although we consider Pinnacle and the Grand Mean Project to be important advances, we believe they are less useful than Propensity Score Select, not least because they constrain the supply of potential respondents for online surveys.

Given the evidence presented here, we believe the Propensity Score Select methodology merits more attention, study, and scrutiny. We are therefore in the process of conducting new
research to evaluate the performance of a live execution of Propensity Score Select rather than this simulation. As part of the research, we will also investigate how different weighting procedures and decisions, in conjunction with Propensity Score Select, affect sample representativeness and response accuracy, among other measures such as cost, time required to field surveys, and the number of potential respondents who need to be screened at the “front door”.

If the Propensity Score Select methodology can deliver on the promise it demonstrated here, then there may be broader implications. By increasing sample representativeness, for instance, it may reduce the risks associated with conducting research with non-probability samples. And by safely opening the door to just about any potential respondent irrespective of his or her originating source, it may accelerate the growth pace of online research in the US and abroad.

It is not difficult to think of other ways in which the Propensity Score Select methodology might benefit the broader research community. For instance, certain respondent selection procedures implemented today in telephone surveys, such as asking to speak only with males after female quotas have been filled, may produce bias, particularly when males who are available for the interview differ from those who are not (and, more generally, from the population of all males). The Propensity Score Select methodology may provide a way to develop more-encompassing quotas. We described the quotas we have in mind—quotas as a set of probabilities—in Section 2.2. If researchers develop such quotas with great care, they might enhance the representativeness of the selected and achieved sample without increasing the costs or time needed to conduct telephone surveys. Exploring the possible benefits and drawbacks of developing and implementing such quotas is also in line with Frank Newport’s call for increased open-mindedness from the AAPOR community.

It is possible, as well, that Propensity Score Select, or similar methodologies, could play a more far-reaching role. Consider, for example, the following scenario:

- The size and character of the accessible population (i.e., the sampling frame) continues to shrink or contort, thereby increasing the threat of non-coverage error.
- The various lists of individuals and telephone numbers—landline, cell phone, business—that are available, when combined, do not cover the entire population, or even a sizeable portion of it.
- People stop answering the telephone.
- Someone—perhaps a “high profile” person such as the son of a senator or a celebrity—is injured or killed while being interviewed via cell phone, with the ensuing outrage leading to a crackdown on unsolicited calls for any purpose.
- Legislators re-conceptualize any telephone research not funded fully by the government as a commercial activity bound by the restrictions of the National Do Not Call Registry.
- The cost of conducting telephone research continues to increase, making it unaffordable to most organizations.

If one or more of these events, some of which are related, were to occur, it would not necessarily doom telephone research (in which respondents are selected through probability sampling), but it would heighten the need to identify a credible alternative. To argue now that Propensity Score Select could swoop in and somehow save the day for telephone research would be premature. We believe it makes good sense, however, for interested parties to investigate that possibility through the conduct of new research, or by providing support for such research. A specific research aim might be to identify the circumstances under which researchers can use Propensity Score Select, or related methodologies, to select telephone respondents who provide information that is representative, valid and affordable.

Many other possible applications for Propensity Score Select come to mind as well. For instance, any researcher who selects respondents from multiple sampling frames (e.g., a landline frame and a cell phone frame) needs to figure out how to pool the resulting responses. Rather than trying to determine in advance precisely how many interviews to complete via each mode, it may make more sense to let the numbers fall where they may based on how the characteristics of the respondents, irrespective of the sampling frame or interviewing mode, compare to those of a benchmark sample. In principle, the steps involved in developing and implementing such a methodology would be similar to those described earlier in Section 5.

Although we are interested in sharing additional ideas, we do not want to put the cart before the horse so we will leave that for another day. As we noted earlier, we will focus our immediate attention on continuing to evaluate and refine the Propensity Score Select methodology for use by (online) market researchers in the US and abroad.
9. References


Appendix. The Questions Used in the Survey (Excluding Non-Demographic Questions Used to Estimate the Propensity Score)

Q0. We’d like to begin by asking you several questions for classification purposes.

Q1. Are you...?
   Male
   Female

Q2. In what state or territory do you currently reside?

Q3. What is your current marital status? Are you . . . ?
   Married
   Living as married
   Single and never been married
   Divorced
   Separated
   Widowed

Q4. Do you (or does your family) own or rent the dwelling in which you live?
   Own
   Rent

Q5. Altogether, including you and any others, how many people regularly live in this household?
   One
   Two
   Three
   Four
   Five
   Six
   Seven
   Eight
   Nine
   Ten or more

Q6. What is your age?
Q7. Are you of Spanish or Hispanic origin, such as Latin American, Mexican, Puerto Rican, or Cuban?

Yes, of Hispanic origin
No, not of Hispanic origin

Q8. Do you consider yourself...?

White
Black
Asian or Pacific Islander
Native American or Alaskan Native
Mixed Race
Some other race

Q9. What was the last grade in school you completed, or degree you received?

8th grade or less
High school incomplete [grades 9, 10, 11]
High school complete [grade 12]
Some college, but no degree
Associate Degree
College Graduate/Bachelors
Postgraduate Degree, such as Master’s, Ph.D., MD, JD

Q10. Which of the following income categories best describes your total 2010 household income before taxes?

Less than $15,000
$15,000 to $24,999
$25,000 to $34,999
$35,000 to $49,999
$50,000 to $74,999
$75,000 to $99,999
$100,000 to $124,999
$125,000 to $149,999
$150,000 to $199,999
$200,000 to $249,999
$250,000 or more
Prefer not to answer

Q11. Would you say your health in general is...?
Excellent
Very Good
Good
Fair
Poor

Q12. Do you approve or disapprove of the way that Barack Obama is handling his job as President?

Approve
Disapprove

Q13. Do you personally own a valid United States passport?

Yes
No

Q14. Do you personally have a valid driver’s license?

Yes
No

Q15. Do you smoke cigarettes...?

Every day
Some days
Not at all

Q16. Which of the following items do you own?

A “smartphone” that can access the Internet
A digital camera
A tablet such as an iPad or Samsung Galaxy
A game console such as Xbox 360, Playstation 3, or Wii
A satellite radio system for your car, home, or other use.
An eBook reader such as Nook, Kindle, or Sony Digital Book

Q17. On the average day, how many hours do you watch television?
Q18. Which of the following have you done online since January 1, 2011?

Made a purchase from an ecommerce website
Banked
Used a social media/social networking application such as Facebook, Twitter, or LinkedIn
Uploaded a picture
 Participated in an online survey other than this one
Watched a video
Participated in an online auction

Q19. To what extent do you consider yourself a religious person? Are you...

Very religious
Moderately religious
Slightly religious
Not religious at all