A smarter way to select respondents for surveys?

George Terhanian and John Bremer
Toluna

Online research has experienced astonishing growth over the past 15 years. To keep up with this growth, researchers have developed new ways of accessing and utilising respondents. Nevertheless, they 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 this reason, it is common today for researchers to use more than one sample source for some types of project, such as a tracking survey that measures change over time. Adding one or more sample source to the original might address the need for more respondents, but some evidence suggests that it might also decrease sample representativeness and reduce response accuracy. In this paper, we introduce a new methodology that enables researchers to select potential survey respondents from either a single sample source or multiple sources based on how well their characteristics match an appropriate, evolving standard with demonstrated evidence of external validity. We also present evidence suggesting that, in the aggregate, respondents who are selected through the new methodology are more representative of the target population than respondents selected by other means. Finally, we consider possible implications of the new methodology on methods other than online research with non-probability samples.

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, researchers know much more about online research, and most interviews (specifically, those commissioned by US market research buyers) are completed online, with US spending expected to exceed $1.8 billion in 2012, or 43% of all survey research spending, and European
A smarter way to select respondents for surveys?

spending expected to exceed €0.7 billion, or 13% of all survey research spending (Inside Research 2012a, 2012b). Part of the attractiveness of online research continues to lie in its low cost (see Duffy et al. 2005 for other reasons for its attractiveness). For instance, the cost today for a completed interview through an online omnibus survey can be as low as $2 in the US. In comparison, the cost for a completed interview through a telephone survey employing a dual frame (landline via random digit dialling and mobile phone via direct dialling) methodology can be at least five times higher.

To keep up with online demand, many market research companies have developed their own opt-in panels (Inside Research 2012a, 2012b). 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 website, 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 overall capacity (i.e. the total number of interviews completed per day).

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 utilising respondents, researchers can still find it difficult to complete the needed number of interviews on time. This is particularly true 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 projects, such as a tracking survey that measures change over time.

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 as the difference between what survey respondents report and what we, or an omniscient observer, know to be true (Bohrnstedt 1983). In research evaluating 17 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 et al. 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, Mitch Eggers has described the Global Market Insights (GMI) Pinnacle methodology
glowingly at industry conferences (see Eggers 2011). At some of those same conferences and in non-peer-reviewed journals (see Gittelman & Trimarchi 2009), Steve Gittelman and Ellen Trimarchi have extolled the virtues of Marketing Inc.’s *Grand Mean Project*. Proponents of these new sample source selection approaches cite at least three benefits: (1) consistency (or interchangeability) of new respondent sources with existing ones; (2) complementariness of new respondent sources with existing ones relative to an external standard; and (3) enhanced representativeness relative to the US general population through calibration with non-online data sources (i.e. external benchmarks such as questions from the US General Social Survey).

Although these new approaches have taken a step in the right direction, we believe they have not gone far enough for at least three reasons: (1) they restrict the pool of potential respondents to those from sample sources vetted previously, thereby limiting supply; (2) they seem to assume that the vetted sample sources do not change over time; and (3) they rely on benchmark data sets that have either limited shelf lives or uncertain external validity. We therefore suspect that they may not produce the same levels of sample representativeness and response accuracy as a new methodology, which we refer to later as *Propensity Score Select*,¹ which selects potential survey respondents, no matter what their originating source might be, based on how well their characteristics match an appropriate, evolving standard with demonstrated evidence of external validity.

We recognise 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, Warren 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).

¹ Commercially, we refer to the methodology as *SmartSelect™*.
More recently, Jon 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 behaviours of a broader population. According to Krosnick, ‘to draw a scientific and representative sample of all residents …, 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 website to refer to sample selection procedures that do not meet the standard Kish put forward as SLOP, an acronym for ‘self-selected opinion polls’. That may also be why some organisations, 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 of the same mind, however. For instance, media organisations such as the Financial Times are less restrictive than the Associated Press and ABC News; they publish the results of such surveys routinely. Many global corporations, including the Walt Disney 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 organisation, attitudes towards online research can differ.

This paper’s aim, given this landscape, is to introduce the Propensity Score Select methodology and, in doing so, describe how and why it may be able to help researchers to select potential survey respondents, irrespective of their originating source (e.g. Panel A, B; River 1, 2), who represent the target population of interest accurately. The latter might include the general population, the online population, past participants in copy testing or brand tracking programmes, or even individuals selected at random to take part in survey research by telephone or face to face.

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.

**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 simple 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, 18 to 24 years old, 25 to 29 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 by gender) and the levels within them (e.g. men, women, 18 to 24 years old), on survey stages (e.g. invitations, starts completes), on sample sources (e.g. the Toluna panel, the YouGov panel), on individual respondent attitudes or beliefs (e.g. attitudes towards technology, political affiliation), or on the many possible combinations among all these factors.

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

---

2 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 capacity.
of all relevant respondent characteristics when using the types of quotas described thus far.

*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 behavioural characteristics, to be a member of the target population rather than an alternative one. As described later here and elsewhere (Terhanian *et al.* 2001; Terhanian 2008), one can estimate each respondent’s predilection, or probability, through the conduct of parallel surveys and the use of logistic regression. Statistical theory, in turn, would help to ensure that individuals who have the same probability also have the same, or a very similar, joint distribution on all variables included in the regression model.

*Historical antecedents of a ‘parallel surveys’ approach*

The idea of using ‘parallel surveys’ to compare and combine (among other possibilities) samples selected by different means (e.g. probability and non-probability sampling) can be traced back nearly 60 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 behaviour of adult males, in their review of the methodology of the Kinsey Report (Kinsey *et al.* 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 *et al.* 1954, p. 23).

Cochran *et al.* 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). 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 state-wide 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 (Rosenbaum & Rubin 1983, 1984) invented a selection bias modelling 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 behaviours, activities or programmes (such as the SAT). To estimate 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).

**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%) 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 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 & Stanley 1963; Cook & Campbell 1979). Campbell and colleagues were also vigilant (e.g. see Campbell & Boruch 1975) to point out the limitations of methods other than random assignment (or ‘randomised experiments’) for that purpose. They warned that such methods do not control for unobserved differences between matched individuals or groups.

**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 & Terhanian 1996, 1998). At the time, Boruch and Terhanian observed that education surveys sponsored by the US federal government rarely explored the individual’s intentions or reasons for membership in a programme or group. They also noted that, at times, researchers would use data from those surveys (in addition to, or instead of, randomised controlled experiments and other techniques) to attempt to estimate the effects of programme 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 programme or group effects.

In practice, applying tools such as propensity scoring to improve survey design can be quite challenging, not least because researchers have
typically used such tools opportunistically to re-analyse existing data sets. That would probably not surprise Leslie Kish, who observed that ‘over 95% 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, 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.

**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). The approach is akin to the one Cochran et al. described in 1954.

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-randomised 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.

In theory, Harris can run into trouble in those later surveys 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 or a similar measure, 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.3 By placing the questions researchers require to estimate the propensity score at the survey’s ‘front door’, they will reduce the risk of including the wrong people, and the related risk of not including enough of the right people.

**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 (e.g. a high-quality telephone or face-to-face survey and an online survey) 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-randomised 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 and several outcome variables. For this illustration, we will regard the US general population as the target population.

---

3 We are not alone in investigating the possible merits of matching methodologies for selecting individuals for surveys. For instance, Rivers (2007, 2012) has described a matching methodology, which he refers to both as ‘sample matching’ and ‘matched sampling’, which the company YouGov has used to apparent good effect to select respondents for US pre-election polls.
It is important to keep in mind that, 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.

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 programme, among many other possibilities.

**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 for a US-based model 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 behaviour, measures of physical activity, and selected attitudes, behaviours and opinions measured on surveys such as the General Social Survey (GSS), 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 et al. 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.

Put somewhat differently, ‘to do a better job constructing propensity scores, one ought to observe or elicit information on why or how people find their way into groups ... with the responses to the question ... incorporated into a propensity score that is better than ... one that relies solely on demographic information’ (Boruch & Terhanian 1996, Ch. 4, p. 29). To that end, Boruch and Terhanian suggested, among other possibilities, cognitive research in a laboratory or field setting, with said
research constituting initial evidence upon which to build theory. Their suggestion is as relevant today as it was then, not least because it can be challenging to figure out how individuals make their way into one group rather than another. We therefore agree, at least partly, 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 regard the latter sentiment more as a starting point 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 behaviour (or, depending on timing, candidate or party preference for an imminent election), possession of a driver’s licence, possession of a passport, and certain buying behaviours. Attitudes, behaviours and opinions measured on surveys such as the GSS, NHIS, NHANES and possibly others.

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 modelling, 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 minimise the differences on the outcome measures (selected in the third step) among control and treatment group members with the same, or similar, propensity scores.
A key output of the exercise would be a distribution of propensity scores for the control group, whatever it may be, that researchers can then use to decide which respondents to select for future surveys.

**Step 5**

Exercise good judgement when deciding, in particular, whether it is possible to represent the attitudes, experiences and behaviours 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 *et al.* 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 Washington, DC, or to explore other matters correlated with online usage, then methods other than online research are probably a better option. In other words, online research is not always fit for purpose.

**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 panellists to minimise 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 favourably 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 dropouts by quintile or decile, implementation problems, time constraints), then weighting might help to reduce bias.

**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 behaviours of target populations will change, so the model development process described here will need to be repeated periodically to account for such change.

**New empirical evidence**

In what follows, we evaluate the effectiveness of the *Propensity Score Select* methodology using data from a 15-minute survey among US adults. The questions in the survey explored various topics, including general issues, attitudes towards privacy, technology ownership and online behaviours. We describe the topics below in more detail. Hereafter, we refer to them as ‘content’ questions.

- **General Issues**: Quality of health, Approval for President Obama, Degree of religiousness, Own passport, Possess driver’s licence, Smoke cigarettes, TV viewing per week
- **Attitudes Towards Privacy**: AIDS screening at work, Unsolicited calls for selling purposes, Cookies on computers for tracking purposes, Airport searches based on visual profiles
- **Technology Ownership**: Smartphone, Digital Camera, Tablet Computer, Game Console, Satellite Radio, eBook Reader
- **Online Behaviours** (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.
We provide below additional details on the four samples.

- **Telephone (Landline):** 1,019 completed interviews among respondents residing in private households in the continental US. Respondents were contacted through random-digit dialling by the Opinion Research Corporation during the period 24–27 February 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):** 1,100 completed interviews during the period 2–6 March 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):** 1,100 completed interviews during the period 2–6 March 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):** 1,098 completed interviews during the period 2–6 March 2011 among members and non-members of Toluna’s online
A smarter way to select respondents for surveys?

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 and household income, as well as the following three measures: (1) willingness to pay higher prices to protect the environment; (2) preference for tailored, personalised communications from companies that are marketing to them; and (3) preference for ‘trying new things’ as soon as they become available.

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 behaviours, the score is less than optimal but more than serviceable based on our experience.

Simulating the effects of the Propensity Score Select methodology

To simulate the effects of the Propensity Score Select methodology, as a precursor to estimating its accuracy, we selected from the combined pool of online respondents a stratified random sample of 1,000 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.

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 US 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 recognise that any of these measures could be biased for various reasons.

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.

‘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.

**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>PSS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region, ‘South’</td>
<td>37.2</td>
<td>36.6</td>
<td>36.6</td>
<td>35.9</td>
<td>35</td>
</tr>
<tr>
<td>Gender, ‘Female’</td>
<td>51.4</td>
<td>50</td>
<td>50</td>
<td>52.4</td>
<td>53.1</td>
</tr>
<tr>
<td>Age grouping, ‘45–54’</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, ‘White, NH’</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, ‘HS degree’</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></th>
<th>Benchmark</th>
<th>Direct invitation</th>
<th>River</th>
<th>Router</th>
<th>PSS*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>1,100</td>
<td>1,100</td>
<td>1,098</td>
<td>1,000</td>
<td></td>
</tr>
<tr>
<td>Mean absolute deviation</td>
<td>1.5</td>
<td>2.8</td>
<td>1.8</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>1.7</td>
<td>8.4</td>
<td>3.1</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td><strong>MSE</strong></td>
<td>4</td>
<td>16.1</td>
<td>6.5</td>
<td>2.2</td>
<td></td>
</tr>
<tr>
<td><strong>MSE vs PSS</strong></td>
<td>–80%</td>
<td>–632%</td>
<td>–194%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*PSS = Propensity Score Select
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 set upper limits on the education level quotas.

‘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.

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. A possible explanation is that Router respondents are more likely than others to be married. An alternative explanation is that married respondents were in high demand in the Router during the field period, and there were simply too few available. Of course, explanations can also have 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 that 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>PSS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household income, ‘$50–$74.99k’</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, ‘Married’</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, ‘Own’</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, ‘2’</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>Benchmark</th>
<th>Direct invitation</th>
<th>River</th>
<th>Router</th>
<th>PSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>1,100</td>
<td>1,100</td>
<td>1,098</td>
<td>1,000</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>

Overall accuracy of all ‘demographic and household’ questions

Table 4 summarises 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 recognise 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>PSS</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>1,100</td>
<td>1,100</td>
<td>1,098</td>
<td>1,000</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>

‘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% 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>PSS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health, ‘Good’</td>
<td>32.3</td>
<td>32</td>
<td>33.6</td>
<td>32.2</td>
<td>31.5</td>
</tr>
<tr>
<td>Obama, ‘Approve’</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, ‘Moderately’</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, ‘Do not own’</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 licence, ‘Have’</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, ‘Not at all’*</td>
<td>80.7</td>
<td>73.5</td>
<td>64</td>
<td>68.6</td>
<td>77.8</td>
</tr>
<tr>
<td>TV, ‘2 hrs per week’</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>1,019</td>
<td>1,100</td>
<td>1,100</td>
<td>1,098</td>
<td>1,000</td>
</tr>
<tr>
<td>Mean absolute deviation</td>
<td></td>
<td>4.5</td>
<td>5.3</td>
<td>4.5</td>
<td>3.4</td>
</tr>
<tr>
<td>Variance</td>
<td></td>
<td>12.7</td>
<td>38.5</td>
<td>22.6</td>
<td>10.6</td>
</tr>
<tr>
<td>MSE</td>
<td></td>
<td>32.6</td>
<td>66.4</td>
<td>43.3</td>
<td>21.8</td>
</tr>
<tr>
<td>MSE vs PSS</td>
<td>−49%</td>
<td>−204%</td>
<td>−98%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The National Household Interview Survey (2010) is the source of the percentage of adults who report they smoke cigarettes ‘Not at all’.*
‘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% 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>PSS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIDS screening, ‘Violation’</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, ‘Violation’</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, ‘Violation’</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, ‘No violation’</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

<table>
<thead>
<tr>
<th></th>
<th>Benchmark</th>
<th>Direct invitation</th>
<th>River</th>
<th>Router</th>
<th>PSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>1,019</td>
<td>1,100</td>
<td>1,100</td>
<td>1,098</td>
<td>1,000</td>
</tr>
<tr>
<td>Mean absolute deviation</td>
<td>5.3</td>
<td>6.9</td>
<td>6.2</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>18.5</td>
<td>17.2</td>
<td>8.7</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>46.4</td>
<td>64.2</td>
<td>47.5</td>
<td>8.9</td>
<td></td>
</tr>
<tr>
<td>MSE vs PSS</td>
<td>−419%</td>
<td>−618%</td>
<td>−432%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

‘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.

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 percentage of adults who own different technology products), appears to contradict advice given by Duffy and colleagues (Duffy et al. 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’ (p. 636).

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.

‘Online behaviours’ questions

Six questions were included in the ‘Online behaviours’ 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% 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 behaviours.

Accuracy of all ‘Content’ questions

Overall, we evaluated 23 questions across the survey’s four main content sections. Table 9 shows that responses from individuals selected through
A smarter way to select respondents for surveys?

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% 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.

**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 more accurate information than 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 may need to spend more time developing and

Table 8  Accuracy of the responses to ‘online behaviours’ (since 1 January 2011)

<table>
<thead>
<tr>
<th>Modal response</th>
<th>Benchmark (%)</th>
<th>Direct invitation (%)</th>
<th>River (%)</th>
<th>Router (%)</th>
<th>PSS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Made online purchase, ‘No’</td>
<td>56.1</td>
<td>50.6</td>
<td>57.3</td>
<td>50.9</td>
<td>57.6</td>
</tr>
<tr>
<td>Banked online, ‘Yes’</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, ‘Yes’</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, ‘Yes’</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, ‘Yes’</td>
<td>70.9</td>
<td>72.5</td>
<td>73.8</td>
<td>75.5</td>
<td>75.4</td>
</tr>
<tr>
<td>Online auction, ‘No’</td>
<td>86.4</td>
<td>73.9</td>
<td>78.5</td>
<td>76.1</td>
<td>83.7</td>
</tr>
</tbody>
</table>

**Key measures**

<table>
<thead>
<tr>
<th></th>
<th>Benchmark</th>
<th>Direct invitation</th>
<th>River</th>
<th>Router</th>
<th>PSS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Base</strong></td>
<td>846</td>
<td>1,100</td>
<td>1,100</td>
<td>1,098</td>
<td>1,000</td>
</tr>
<tr>
<td><strong>Mean absolute deviation</strong></td>
<td>7.8</td>
<td>5.6</td>
<td>6.9</td>
<td>5.1</td>
<td></td>
</tr>
<tr>
<td><strong>Variance</strong></td>
<td>27</td>
<td>8.2</td>
<td>18.8</td>
<td>10.9</td>
<td></td>
</tr>
<tr>
<td><strong>MSE</strong></td>
<td>88.4</td>
<td>39.4</td>
<td>66</td>
<td>36.7</td>
<td></td>
</tr>
<tr>
<td><strong>MSE vs PSS</strong></td>
<td>–141%</td>
<td>–7%</td>
<td>–80%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
improving respondent selection procedures and less time tinkering with source selection ones. 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.

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 has demonstrated here, then there may be broader implications. By increasing sample representativeness and response accuracy, for instance, it may further reduce the risks associated with conducting research with non-probability samples. And, by opening the door safely to just about any

<table>
<thead>
<tr>
<th>Key measures</th>
<th>Benchmark</th>
<th>Direct invitation</th>
<th>River</th>
<th>Router</th>
<th>PSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>1,019</td>
<td>1,100</td>
<td>1,100</td>
<td>1,098</td>
<td>1,000</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>

Table 9 Accuracy of the responses to all ‘content’ questions
potential respondent irrespective of his or her originating source, it may accelerate the growth pace of online research.

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; specifically, quotas as a set of probabilities representing the predilection of each potential survey respondent, given his or her demographic, attitudinal and behavioural characteristics, to be a member of the target population rather than an alternative one.

It is possible, as well, that Propensity Score Select, or similar methodologies, could affect telephone research in an even more significant way. Consider, for example, the following scenario.

- The percentage of households with landlines continues to decrease.
- The available listings of households, individuals and telephone numbers (e.g. landline, mobile, business), when combined, do not cover the entire population, or even a sizeable portion of it.
- People answer the telephone less frequently, if at all.
- Someone – perhaps a ‘high profile’ person such as the son of a Senator, the daughter of a Member of Parliament, or a celebrity – is injured or killed while being interviewed via mobile phone, with the ensuing outrage leading to a crackdown on unsolicited calls for any purpose.
- Legislators re-conceptualise any telephone research not funded fully by the government as a commercial activity bound by the restrictions of the US National Do Not Call Registry, or similar registries in other countries.
- The cost of conducting telephone research continues to increase, making it unaffordable to most organisations.

If this kind of scenario 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 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 mobile 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 or dataset. In principle, the steps involved in developing and implementing such a methodology would be similar to those described earlier.

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 noted earlier, we will focus our immediate attention on evaluating, refining and expanding the Propensity Score Select methodology for use by online market researchers in the US and elsewhere.

Appendix: The questions used in the survey

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?
A smarter way to select respondents for surveys?

- 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, PhD, 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
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 licence?
- 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 smarter way to select respondents for surveys?

- 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 1 January 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

Q20. How willing would you be to pay much higher prices in order to protect the environment? Would you say you are …?
- Very willing
- Fairly willing
- Neither willing nor unwilling
- Not very willing
- Not willing at all

Q21. Many companies today want to know as much as they can about their customers’ interests, needs and preferences so they can individually tailor their communications and offers to each person. In general, do you see such methods for improved targeting as a good thing?
- Yes, it is a good thing
- No, it is not a good thing

Q22. Which of the following statements describes you best?
- I rarely like to try new things
❑ I like to try new things only after they have been in the marketplace for a while
❑ I like to try new things but only after other people have tried them and recommended them to me
❑ I like to try new things as soon as they come out

References


A smarter way to select respondents for surveys?


---

**About the authors**

George Terhanian is Toluna’s North American President and Group Chief Strategy Officer. He is a member of the boards of directors of the Advertising Research Foundation and the Council of American Survey Research Organizations as well. Before joining Toluna in 2010, Dr Terhanian served in a variety of roles for Harris Interactive, including President, Global Solutions; President, Harris Interactive Europe; and President, Global Internet Research. He also oversaw the Harris Poll. His methodological expertise lies in the design and analysis of multi-mode studies.

John Bremer is Toluna’s Chief Research Officer. His work has consisted of research into all areas of representativeness, specialising in areas of non-probability sampling, rare populations, selection bias, and online survey research. Prior to joining Toluna in 2011, he co-directed the Harris Center for Innovation and played a lead role in producing successful election forecasts in the US and UK.

Address correspondence to: George Terhanian, Toluna, 21 River Road, Wilton, Connecticut, 06897, USA.

Email: george.terhanian@toluna.com