Survey Sampling Methodology

Loren Collingwood, Ph.D.*

April 17, 2014

Survey research is presently – and has been – undergoing major changes in data collection. Fewer respondents permit interviews these days compared to even five and 10 years ago. Further, landline telephone use is on the decline. Large percentages of younger citizens do not even have access to a landline phone. At the same time, internet usage is steadily increasing among all demographic subgroups – although representation challenges still remain. ClearPath Strategies has developed an innovative sampling strategy blending data collected from online internet panels with cell and landline interviews. This document discusses the main challenges facing telephone survey methodology and the growing promise of online data collection. In addition, the document reviews the growing literature on propensity score matching, and overviews ClearPath's general sampling and data collection strategy.

1 INTRODUCTION

In the last several decades, survey sampling and data collection in the United States have undergone tremendous changes due mainly to technological changes. Initially, survey data were collected via face-to-face cluster sampling techniques (Kish, 1965), but with the rise of telephone universality the bulk of data collection shifted to the telephone. Then again, following the internet boom, many surveys are now collected online. Each approach has benefits and downfalls, but the trend is clear: in the face of declining telephone response rates

* Collingwood is Assistant Professor of Political Science at University of California, Riverside. He is a former Senior Researcher at the Washington Poll, an academic polling institute at the University of Washington, and an analyst at Greenberg Quinlan Rosner, a private polling outfit based in of Washington, D.C.

and cost, internet survey sampling is the wave of the future. ClearPath Strategies employs a variety of sampling techniques, which this document outlines.

First, in the face of declining telephone response rates, this document briefly examines the literature on survey non-response and statistical inference. Second, a literature review on a relatively new sample design, matching, is discussed. The final component of this document outlines ClearPath's sampling approach, which includes the use of landline and cell phone samples, as well as matched online panel data. All data are weighted to established and known population estimates.

2 Probability-Based Samples and Non-Response

Telephone and face-to-face random samples are typically probability-based; that is, each respondent in the target population has an equal probability of selection (EPSEM). Probability sampling is usually done in two ways. One common technique is based on listed sampling: researchers gather a list of respondents/households from the sampling frame and then employ some sort of random draw from the list. The other common approach with telephone sampling is to employ a random digit dial technique (RDD). Here, telephone area codes and exchanges are selected proportionate to size and the other numbers randomly spun (Groves et al., 2001). In both cases, a random sample is developed, and thus statistical inferences – such as a calculation of the statistical margin of error – to the target population can be made (Cochran, 2007).

However, the theory behind probability sampling often faces validity challenges due to declining response rates in telephone surveys. That is, increasingly only a very small percentage of contacted telephone respondents willingly take any given survey. Non-response is not necessarily a statistical problem if survey responders are no different from non-responders on survey items of interest (Groves, 2006). However, if responders differ from non-responders on key demographics and attitudinal questions, then statistical bias is present in the data and must be corrected via weighting. Further, as response rates have declined, researchers have increasingly relied on post-survey weighting to produce population estimates, which, after a point, can become unstable and actually increase the margin of error due to design effects (Groves et al., 2013). That is, for example, if a survey has relatively few young African-Americans relative to their overall population, their weight value may be upwards of five or six. The survey is thus unlikely to capture much variation within this group. Many survey researchers have thus begun investigating other avenues of data collection.

Table 2.1 shows the change in contact (percent of households where an adult was reached), cooperation (percent of contacted households that produced an interview), and response rates (percent of sampled households that produced an interview) across time for Pew Research U.S. surveys. Between 1997 - 2012, the response rate declined by 27 points from 36% to just 9%. This is a marked drop facilitated by changing telephone and communication technology, and presents a massive problem to survey responders. Pew – a very transparent and reliable polling outfit – surveyed their data for five days, which is a relatively typical time period for many polling and consulting outfits. While it is impossible to know for sure the response rate for private polling organizations, many collect their data in just three day blocks, and almost certainly the response rates are below 9%.

	1997	2000	2003	2006	2009	2012
Contact Rate	90	77	79	73	72	62
Cooperation Rate	43	40	34	31	21	14
Response Rate	36	28	25	21	15	9

Table 2.1: Pew Research Center 2012 Methodology Study

In the report, *Assessing the Representativeness of Public Opinion Surveys*, Pew finds that differences on attitudinal and demographic variables between high (government survey with 75% response rate) and low response rates (Pew, January 2012) are relatively small although some differences do exist on many participatory questions indicating that typical survey respondents are much more likely to be engaged in politics than are non-respondents (Kohut et al., 2012).[1] While post-stratification weighting can rectify many of the problems associated with increasing non-response rates, in the context where private polling firms likely have response rates below 5%, EPSEM principles are unmistakenly questionable. Given this scenario and increasingly expensive nature of phone interviewing, many researchers have investigated the possibility of using opt-in online panels for data collection needs.

3 Online Panels and Matching

ClearPath Strategies employs a mixed mode survey design including landline phone, mobile phone, and opt-in internet panel samples. This design allows ClearPath to reach populations that are less likely to be online (namely, seniors, some minority groups, and the poor) and less likely to use landline phones (namely, younger respondents), therefore getting around the issue of low landline response rates among youth and other subgroups. While both the landline and mobile subsamples are probability-based, the opt-in internet sample is non-probability based. This section briefly reviews opt-in panels, the matching literature, and describes the matching procedure typically used to correct for opt-in panel sample bias.

There are two main problems with using opt-in online panels. The first is that online usage is not as wide as telephone (land and mobile usage). However, over time, an increasing number of Americans have gained online access. Indeed, by May, 2013, 85% of U.S. adults reported using the internet at least occasionally (Zickuhr, 2013). Thus, the initial coverage area argument that few people even have a chance to be selected into a sample because few people regularly use the internet is no longer an especially valid argument.

The second issue is that opt-in panels, are, by definition, non-probabilistic. Usually respondents will see a banner ad on a certain website, click on it, and sign onto a panel. The tradeoff for their time is the chance to win prizes and cash. Thus, while these samples can be very accurate, theoretically, statistical inference is not possible because we do not know

[1] Differences on “food stamps or nutrition assistance,” “contacted a public official in past year,” “volunteered for an organization in past year,” “voted in 2010.”

each person in the target population's probability of selection into the sample. Thus, despite the fact that some opt-in web panels have been shown to be quite representative of the target population (Rivers, 2007), calculating statistics such as margin or error and confidence intervals is inappropriate. This is because surveys based on opt-in or self-selection have no known relationship to the target population (Baker et al., 2010).

However, propensity score matching among opt-in internet panels has been shown to approximate RDD samples in terms of demographics and polling accuracy. Rivers and Bailey (2009) assess the performance of YouGov matched sampling during the 2008 presidential election. The researchers find that the matched sample produces little or no bias relative to RDD or other probability-based internet polls (Knowledge Networks) conditional on a survey's demographic controls (base: age, race, gender, education, region; augmented: plus employment status, income, and marital status), and the size of the opt-in panel. For instance, compared to RDD national telephone polls, matched samples produce similar proportions of respondents in the following categories: race, education, gender, marital status, and region. Moreover, compared to the telephone samples, the matched sample produces comparable univariate and bivariate 2008 Obama vote estimates.

In the 2006 midterm elections, Vavreck and Rivers (2008) analyzed the accuracy and bias of the 2006 Cooperative Congressional Election Study employing a similar approach as Rivers and Bailey (2009). They show that results outperform RDD results from a similar time period, and conclude that opt-in internet panels are a cost-effective alternative to traditional sampling and survey data collection. Thus, while statistical properties do not technically apply to opt-in panels, much evidence indicates that survey results produced from these panels are indeed quite accurate relative to other modes of data collection.

4 ClearPath Sampling

As noted, ClearPath uses a propensity score matching process to develop its online sample. In the context of a statewide survey, first, ClearPath generates a random sample of a specified number of respondents (e.g., 750 respondents) from a state's (e.g., California) voter file. Working with well known and experienced sample vendors, the voter files come augmented with many variables, such as race, income, and estimated education. In the parlance of matching, the voter file is treated as the control group, which is the data that is matched against the online panel. The treatment group is therefore the opt-in panel. Using a distance function – in this case a genetic algorithm – ClearPath matches each person from the control group against someone from the online panel (Ho et al., 2007; Sekhon, 2011). The matching process takes into account the following variables: gender, age, race, education, income, marital status, and region. For example, if someone in the control (i.e., random) sample is Latino, age 22, has a college degree, makes $45,000, is single, and lives in Los Angeles, the matching algorithm searches for the person that best fits these characteristics. This matching is possible because voter file data are merged onto the opt-in panel so the same variables are in both datasets.

Finally, because online panel respondents tend to under-represent important populations (e.g., voters over the age of 70, some minority populations, and urban and rural poor), ClearPath balances the online data with landline and cell phone interviews. To ensure the data are representative of the overall population and to account for differences in cellular and landline use by subgroup, ClearPath further sets subgroup quotas on region, cellular and landline use, and online usage. The final result is a balanced and representative sample on age, region, race, gender, and education – all before the weighting process. Any final post-survey imbalances are corrected via post-stratification weighting based on counts from an estimated likely voter model of pooled 2013 surveys from Public Policy Institute of California (PPIC).

5 Summary

In conclusion, survey researchers – especially market, commercial, and media pollsters – operate in a much different world today than even 10 years ago. Declining response rates have increased the cost of survey research, reliance on weighting, and even call into question notions of probability-based sampling. To this end, many researchers have begun turning to opt-in online panels for data collection. Opt-in online panel surveys present their own challenges, but those can be overcome by taking the right steps, such as those outlined here. While not all pollsters use a matching algorithm to correct for opt-in sample bias, the research suggests matching is the appropriate method to handle non-probabilistic data. Therefore, ClearPath's methodology employing propensity-score matching online interviewing in conjunction with traditional landline and cell phone interviewing, is – collectively – both a representative and relatively affordable solution on a mass scale.

References

Baker, R., S. J. Blumberg, J. M. Brick, M. P. Couper, M. Courtright, J. M. Dennis, D. Dillman, M. R. Frankel, P. Garland, R. M. Groves, et al. (2010). Research synthesis aapor report on online panels. *Public Opinion Quarterly 74*(4), 711–781.

Cochran, W. G. (2007). *Sampling techniques.* John Wiley & Sons.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly 70*(5), 646–675.

Groves, R. M., P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls, J. Waksberg, et al. (2001). *Telephone survey methodology*, Volume 328. John Wiley & Sons.

Groves, R. M., F. J. Fowler Jr, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2013). *Survey methodology.* John Wiley & Sons.

Ho, D. E., K. Imai, G. King, and E. A. Stuart (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis 15*(3), 199–236.

Kish, L. (1965). Survey sampling.

Kohut, A., S. Keeter, C. Doherty, M. Dimock, and L. Christian (2012). Assessing the representativeness of public opinion surveys. *Pew Research Center, Washington, DC.*

Rivers, D. (2007). Sampling for web surveys. In *Joint Statistical Meetings*.

Rivers, D. and D. Bailey (2009). Inference from matched samples in the 2008 us national elections. In *Proceedings of the Joint Statistical Meetings*, pp. 627–639.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: the matching package for r. *Journal of Statistical Software 42*(7), 1–52.

Vavreck, L. and D. Rivers (2008). The 2006 cooperative congressional election study. *Journal of Elections, Public Opinion and Parties 18*(4), 355–366.

Zickuhr, K. (2013). *Who's not online: 15% of American adults do not use the internet at all, and another 9% of adults use the internet but not at home.* Pew Internet & American Life Project.