Marketing Systems Group

SCIENTIFIC SURVEYS BASED ON INCOMPLETE SAMPLING FRAMES AND HIGH RATES OF NONRESPONSE

Fahimi, M., F. M. Barlas, R. K. Thomas and N. Buttermore. 2015. Scientific Surveys Based on Incomplete Sampling Frames and High Rates of Nonresponse. *Survey Practice.* 8 (5).

Abstract

Traditional methods of survey research that rely on Neyman's probability-based sampling paradigm are grounded in a number of fundamental assumptions that are becoming exceedingly difficult to attain in today's survey research environment. On the one hand, common methods of sampling are subject to coverage issues that may not be fully ameliorated through post-survey weighting adjustments. On the other, response rates continue to deteriorate for all surveys, even when resource-intensive refusal conversion strategies are employed. Add in the growing need for cost containments and it is no wonder why alternative sampling methods are gaining popularity. The authors will review a number of practices that are currently used for developing inferences from samples that do not fully adhere to the statistical machinery that is currently available for probability-based sample surveys. Moreover, a robust weighting methodology will be introduced that can reduce the inherent biases associated with non- probability samples, as well as probability-based sample surveys that suffer from incomplete frames and high rates of nonresponse. The efficacy of the proposed methodology is assessed in light of comparisons of survey estimates to external benchmarks, relying on parallel surveys that were conducted in two states using both probability-based and non-probability samples.

Overview

For decades, the traditional methods of probability-based sampling have served as the gold standard for survey research applications. Relying on the statistical machinery developed by Neyman (1934), it has been possible to make measurable inferences about target populations, when sampling units carry known selection probabilities and samples are selected from complete sampling frames. Moreover, any observed non-response has been explained away either by assuming randomness of non-response or by applying compensatory adjustments when differential non-response has been deemed non-ignorable. Grounded in this solid framework and relying on fairly high rates of response, well designed and executed surveys have been able to produce reliable estimates of population parameters based on relatively small samples. As such, sample surveys have served as a foundation for data-driven decision-making processes.

However, the two main tenets of survey sampling – availability of complete sampling frames and high rates of response – are becoming exceedingly difficult to secure because many surveys are subject to growing coverage problems and eroding rates of response (Biener et al. 2004; Keeter et al. 2006). Consequently, survey researchers are forced to rely more heavily on geodemographic weighting adjustments to compensate for undercoverage and non-response. Such bias reduction, however, comes at the expense of diminished precision of surveys because weighting inflates variance of survey estimates (Fahimi et al. 2007).

It is in this context, when many surveys have to settle for low response rates and sampling frames with varying levels of undercoverage, that probability- based sample surveys are beginning to lose their bragging rights as compared to less expensive alternatives that employ convenience sampling methods. After all, there is only so much traditional weighting adjustments can accomplish in realigning survey respondents to represent their target populations – even when tolerating significant blows to the precision of survey estimates due to unequal weighting effects. In light of such formidable challenges, it has been suggested the future of sampling is likely to be in the hands of personalities who have not yet been revealed (Brick 2011).

Need for Innovation

Unlike a century ago when full enumeration was deemed the only reliable method for population studies, in recent decades probability-based survey sampling has emerged as a universally accepted alternative for creating reliable and cost-effective population statistics (Kruskal and Mosteller 1980). With the main pillars of this methodology – availability of complete sampling frames and high rates of response – beginning to crumble, it can be argued that survey sampling will be best served if researchers adopt a two-pronged approach when investigating innovative options for the future.

First, it is crucial to develop cost-effective methods of sampling and survey administration options that can improve coverage while reducing non-response at the same time. Recent improvements in address-based sampling (ABS) and dual-frame RDD (DFRDD) methodologies are examples of this line of investigation (Fahimi 2009a, 2009b). Moreover, other sample survey protocols have to be considered that are not rooted in the classical probability-based paradigm. Increased tolerance for such alternatives is inevitable, since sample surveys based on the traditional methods – in addition to coverage and response rate issues – are often cost-prohibitive for many applications (Baker et al. 2013).

Second, more effective remedial measures have to be investigated that can compensate for the growing rates of undercoverage and non-response. This is true for both probability-based and non-probability samples, since both samples can end up misrepresenting their target populations in measurable and unmeasurable ways. It is from this perspective that this paper examines a weighting (calibration) methodology that goes beyond commonly used geodemographic weighting adjustments since, in many instances, such geodemographic

adjustments no longer provide adequate corrections in the presence of severe undercoverage and non-response.

Calibration 1.0

KnowledgePanel (KP, GfK North America, New York, NY, USA) is the largest online panel in the United States with over 55,000 members for which panelists are selected with known probabilities from an ABS frame that represents US households. However, most online surveys depend on samples that are comprised of non-probability samples, including unknown groups of online users who have opted to join such panels for ad-hoc survey participation. DiSogra et al. (2011) proposed a methodology whereby a probability-based KP sample is supplemented with one that is non-probability from opt-in (OP) panels to increase the combined sample size or deal with surveys of hard-to-reach subgroups. Specifically, an identical online instrument is used to administer surveys to samples selected from the KP and OP panels. DiSogra et al. demonstrated that OP respondents, as compared to those from KP, tend to score significantly higher on a short battery of questions that measure early adoption (EA) of new products and services:

EA1. I usually try new products before other people do

EA2. I often try new brands because I like variety and get bored with the same old thing

EA3. When I shop I look for what is new

EA4. I like to be the first among my friends and family to try something new

EA5. I like to tell others about new brands or technology

Armed with the above observable differences between OP and KP respondents, a calibration weighting adjustment was developed that attempts to correct for the systematic bias due to the higher propensity of OP respondents to be early adopters. This methodology is rooted in techniques described by Skinner (1999) and Kott (2006) in which the needed calibration benchmarks are obtained from the parallel online probability KP survey that is separately weighted to standard geodemographic benchmarks. Subsequently, the combined calibrated OP and KP data produce survey estimates that not only match the EA distributions –enforced by calibration – but also exhibit improved internal validity with respect to other population parameters as estimated by the weighted KP data.

A slightly refined version of the above methodology combines the KP and OP surveys using an optimal blending process that is based on their respective effective sample sizes (Fahimi 1994). Specifically, once the KP component has been weighted to the standard geodemographic benchmarks, study specific distributions of the EA battery are generated for calibration of the OP component. Subsequently, the OP sample component is weighted to the same standard geodemographic benchmarks as well as the KP-based EA distributions. Next,

the effective sample sizes for the two components are computed, for which the design effect for KP component only accounts for the unequal weighting effect due to poststratification. This is necessary because for OP samples there are no design weights available, and hence, their "design effects" only reflect the final poststratification – quota-driven samples can yield exceptionally low pseudo design effects. In the final step, the two components are blended in proportions of their respective effective sample sizes and reweighted one last time to the combination of the geodemographics and EA distribution benchmarks.

While the above simple adjustment carries an intuitive and pragmatic appeal, EA attributes are not the only measures with respect to which OP and KP respondents differ significantly. Moreover, a series of factor analyses revealed that all of the above five EA attributes tap into the same latent measure, rendering the proposed method a univariate calibration adjustment. As outlined in the next section, there are other measures which indicate that the two pools of respondents think and behave differently. This research seeks to identify a set of core differences and develop a multivariate calibration adjustment methodology that improves not only the internal validity, but also external validity of survey estimates. This methodology is applicable to both surveys that rely on probability-based samples that are subject to high rates of undercoverage and non-response, as well as OP samples selected with unknown probabilities.

Calibration 2.0

TIn order to improve the existing calibration methodology and evolve it from a univariate procedure to one that is more comprehensive and multivariate in nature, a number of parallel assessments were carried out. First, it was necessary to identify other behavioral and attitudinal dimensions that could effectively differentiate between the two pools of respondents. This task was accomplished by conducting several KP and OP surveys in parallel that included a common set of questions on a diverse set of topics. These questions were secured from a series of brainstorming sessions with subject matter experts, as well as other research streams dealing with reducing bias for non- probability samples, including: Duffy et al. (2005); Lee (2006); Rainie et al. (2013); Schonlau et al. (2007); Smith et al. (2013); Terhanian and Bremer (2012).

Once survey data were collected on this long list of potential differentiators, the emerging list was winnowed down to only a few dozen questions about which KP and OP respondents had provided significantly different responses in a number of parallel studies. These differences were detected after both sets of data were weighted to a comprehensive set of geodemographic variables to eliminate confounding effects. A brief listing of the emerging differentiators and their underlying themes are listed in Table 1.

In the third step, additional parallel surveys were conducted to identify which subset of the resulting significant differentiators could serve as new calibration variables to improve the external validity of the survey results. In addition to including a top list of differentiating questions, an ancillary list of questions were seeded in the survey instruments for the

objective of estimating population parameters for which reliable external estimates were available (see Tables 2 and 3). For each estimate, its corresponding mean squared error (MSE) was computed by reflecting its design-proper measure of variance and bias as compared to the presumed unbiased estimate obtained from the following government sources:

Moreover, similar comparisons were carried out with respect to a series of election-related measures for which external estimates were available.

Table 1 Significant differentiators between KP and OP respondents.

A. Social engagement:	**F. Community:**
1. Taking vacation with others	1. Feeling part of the community
2. Exercising/playing sports with others	2. Moves in past five years
3. Having meals with others	3. Extent of religiosity
B. Self-assertion:	**G. Altruism:**
1. Importance of opinion sharing	1. Donating blood
2. Strength of opinions	2. Donating items
3. Confidence in social settings	3. Volunteering without pay
C. Shopping habits:	**H. Survey participations:**
1. Using coupons for shopping	1. Experience with online surveys
2. Enjoying shopping online	2. Important of taking surveys
3. Rating brand more important than price	3. Frequency of online surveys
D. Happiness and security:	**I. Internet and social media:**
1. Happiness with life	1. Frequency of personal emails
2. Feeling insecure and lonely	2. Frequency of accessing Internet
3. Concerned about cyber security	3. Time spent watching TV per day
E. Politics:	
1. Having influence on politics	
2. Government's effectiveness	
3. Closely following the news	

Table 2 Ancillary questions for assessing the efficacy of calibration models form government statistics.

A. BRFSS (2013)	**B. CPS (2014)**
1. Smoking 100 cigarettes in lifetime	1. Receiving Social Security
2. Physical check-up in past year	2. Marital status
3. History of depressive disorder	3. Homeownership status
4. Days per month physical health not good	4. Household income less than $25,000
5. Hours of sleep per night	
C. NSDUH (2012–2013):	**D. ACS (2011–2013)**
1. Wearing seatbelt as front passenger	1. Number of bedrooms in house
2. Risk of smoking one or more packs a day	2. Number of automobiles
3. Risk when trying heroin once or twice	

Methodology

As part of a larger study on election outcomes in two states in 2014, two sets of parallel surveys were conducted, one in Illinois and a second in Georgia, using both KP and OP sample components. Both surveys included identical batteries of questions that could be used for three purposes: standard geodemographic questions for weighting, differentiating questions for experimenting with various calibration models, and ancillary questions for assessing the efficacy of the employed models relative to available external benchmarks. Table 4 provides a summary of the respondent counts for each survey and state. Accordingly, a total of 4,982 surveys were completed; 2,213 in Georgia; and the remaining 2,769 in Illinois. The number of OP surveys completed was nearly twice as large as those for KP surveys.

Using various subsets of the new calibration variables as summarized in Table 1, different calibration models were assessed with the goal of improving the external validity of the survey results. These estimates were produced under each of the following survey scenarios:

A. KP only survey data weighted to the standard geodemographic variables;

B. Combination of the KP and OP survey data blended and weighted to the standard geodemographic variables, with the OP data calibrated using the EA battery (Calibration 1.0);

C. OP only survey data weighted to the standard geodemographic variables;

D. OP only survey data weighted to the standard geodemographic variables and calibrated using the EA battery (Calibration 1.0);

E. OP only survey data weighted to the standard geodemographic variables and calibrated using the new battery (Calibration 2.0); and

F. Combination of the KP and OP survey data blended and weighted to the standard geodemographic variables, with the OP data calibrated using the new battery (Calibration 2.0).

Table 3 Ancillary questions for assessing the efficacy of calibration models from election statistics.

A. Illinois:	**B. Georgia:**
1. Percent registered voters	4. Percent registered voters
2. Percent Republican	5. Percent Republican
3. Percent Conservative	6. Percent Conservative
4. Senate race	7. November Senate race
5. Governor's race	8. November Governor's race
6. Insurance coverage for birth control	

Table 4 Summary respondent counts by survey type and state.

	Number of Respondents		
State	**KP**	**OP**	**Total**
Georgia	654	1,559	**2,213**
Illinois	1,017	1,752	**2,769**
Total	**1,671**	**3,311**	**4,982**

Results

After examining the reduction in average MSE for estimating the government statistics summarized in Table 2, the subset of calibration variables resulting in the largest reduction with the smallest variance inflation was identified as the current Calibration 2.0 model. Given the pragmatic limitations that for most commercial research no more than 6 to 8 questions could be added exclusively for calibration purposes, the identification of this parsimonious subset was achieved in two steps. It should be noted that in the interest of brevity, descriptions of these steps, which included many computational details, are kept to a minimum by highlighting only the key points.

In the first step, a series of CHAID analyses were conducted to identify variables with the highest relative importance for differentiating between KP and OP respondents. Among the emerging top differentiators, in the second step, average MSE for various calibration models with subsets of 6 to 8 variables were computed. Ultimately, the one subset with the smallest overall unequal weighting effect and average MSE was selected as the optimal subset (model). As such, our Calibration 2.0 model included the following subset of variables:

1. Number of online surveys taken in a month;
2. Hours spent on the Internet in a week for personal needs;
3. Interest in trying new products before other people do;
4. Time spent watching television in a day;
5. Using coupons when shopping; and
6. Number of relocations in the past 5 years.

As seen from Figure 1, the external validity of both KP and KP+OP survey data improve significantly when the one-dimensional EA-based calibration (1.0) adjustment is replaced by a multidimensional calibration using the 6 variables listed above (2.0). This validity is measured in terms of estimating statistics reported by the government surveys – BRFSS, NSDUH, CPS, and ACS – that were not controlled for during the weighting/calibration process. A key point to note is that the new calibration model outperforms Calibration 1.0 approach even when the OP survey data are used without any contribution from a KP sample component. Moreover, the KP survey data – alone or blended with OP – provide a higher level of external validity in both states.

Similar assessments were carried out with respect to the variables listed in Table 1 corresponding with the November 2014 election results in Georgia and Illinois. As seen in Figure 2, the same set of conclusions can be drawn when examining such results. That is, the new calibration methodology improves the external validity of survey estimates when considering the blended KP and OP results, as well as when OP data are used for estimation alone.

Conclusions

The survey research industry is currently in a state of flux due to formidable challenges that question the external validity of the statistical machinery we have relied on for decades to develop measurable inferences for population parameters using probability-based samples. Top among such challenges are coverage issues that existing sampling frames are subject to, even those that employ ABS or DFRDD methodologies. Perhaps a more imposing challenge has to do with the deteriorating rates of response to virtually all surveys, even large- scale government surveys. Naturally, these challenges are more pronounced for small-scale and commercial surveys that are constrained by lower budgets and shorter field periods. For example, most online surveys struggle to secure response rates that are higher than single digits, and those based on opt-in samples have completion rates that fall even below one percent.

While in recent decades effective remedies have been developed to deal with coverage and non-response problems, the efficacy of such treatments have come under serious questioning as the magnitude of undercoverage and non-response problems continues to grow. It is one thing to explain away a 10 percent non- response rate for a sample selected from a near-perfect frame by assuming randomness of non-response and perhaps applying some form of non-response adjustment, but it is quite another thing to resort to the same explanations when over 100 percent of sampled units remain unaccounted for. As such, our traditional methods of weighting that rely on basic geodemographic adjustments are becoming increasingly ineffectual.

Our proposed methodology attempts to go beyond traditional weighting procedures by applying more comprehensive adjustments, using behavioral and attitudinal measures that historically have remained outside of consideration for survey weight calculations. Our research, while not claiming to have found the “secrete sauce” for all calibration applications, has shown great promise for reducing systematic biases in today’s survey data. The proposed methodology is applicable not only to non-probability samples, but also probability-based samples from incomplete frames that are subject to high rates of non-response. The efficacy of our methodology is measured with respect to improved inferential properties of calibrated data when estimating population parameters for which high quality estimates are available. These include estimates from government surveys, such as CPS, ACS, BRFSS, and NSDUH, as well as ad-hoc estimates, such as those related to election outcomes.

Finally, we would be remiss not to mention a potential consequence associated with our proposed calibration methodology. Given that bias reduction through weighting is always exercised at the expense of variance inflation, as more variables are included in the weighting/calibration process the smaller the effective sample size of a survey becomes. Perhaps a fitting analogy from the field of medicine in this context would be that, as the severity of an illness goes up the dosage of the required medicine goes up in tandem. Analogously, as the misrepresentation of a sample becomes more severe, stronger weighting adjustments become necessary to recover the health (representation) of the sample in

question. Ignoring this reality in the interest of declaring a larger effective sample size would be wishful thinking, an imprudent practice that can lead to erroneous conclusions based on biased estimates and underestimated error margins with costly implications.

Figure 1 Average MSE estimating government statistics under different weighting/ calibration adjustments by state.

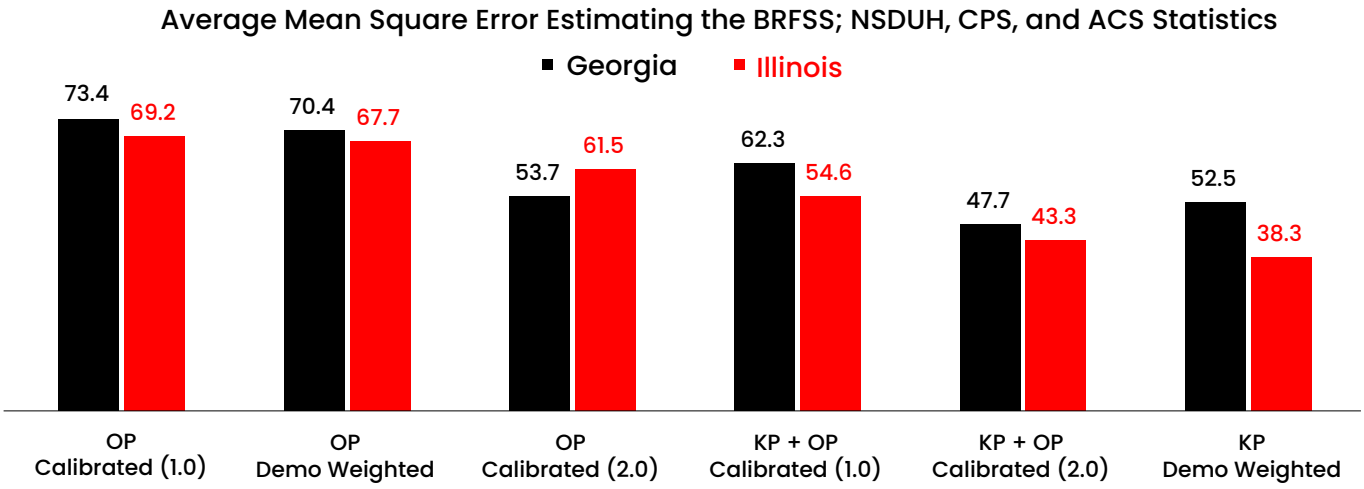
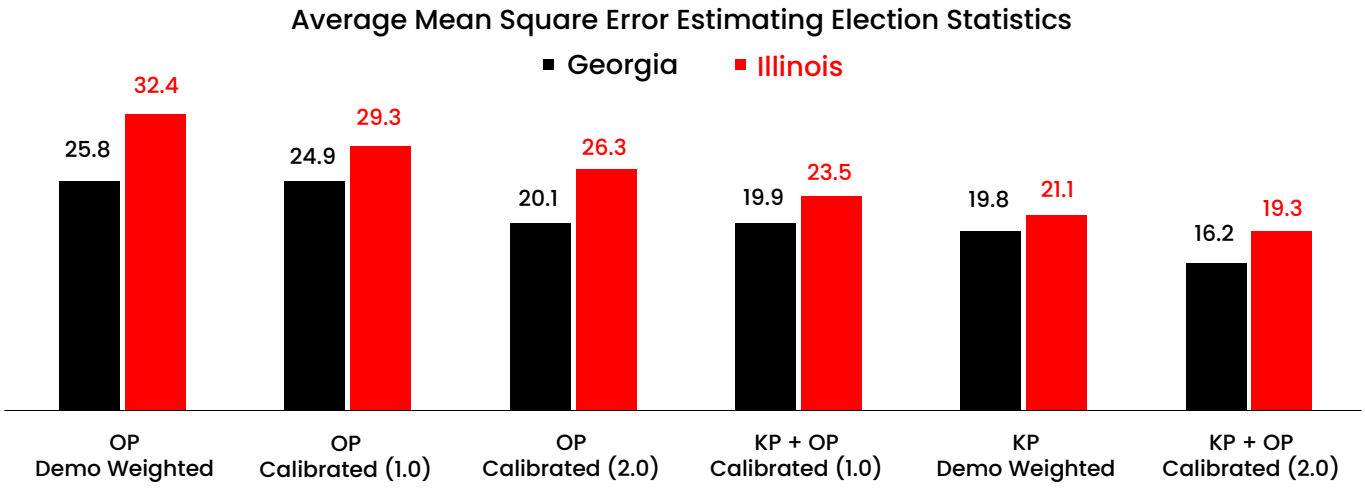


Figure 2 Average MSE estimating election statistics under different weighting/calibration adjustments by state.



References

Baker, R., J.M. Brick, N.A. Bates, M. Battaglia, M.P. Couper, J.A. Dever, K.J. Gile and R. Tourangeau. 2013. The AAPOR Report on Non-Probability Sampling. Available at: http://www.aapor.org/AAPORKentico/AAPOR_ Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13. pdf.

Biener, L., C.A. Garrett, E.A. Gilpin, A.M. Roman and D.B. Currivan. 2004. Consequences of declining survey response rates for smoking prevalence estimates. *American Journal of Preventive Medicine 27*(3): 254–257.

Brick, J.M. 2011. The future of survey sampling. *Public Opinion Quarterly*, Special Issue 75(5): 872–888.

DiSogra, C., C. Cobb, M. Dennis and E. Chan. 2011. Calibrating non- probability Internet samples with probability samples using early adopter characteristics. *Proceedings of the American Statistical Association, Section on Survey Research.* Joint Statistical Meetings (JSM). Miami Beach, FL.

Duffy, B., K. Smith, G. Terhanian and J. Bremer. 2005. Comparing data from online and face-to-face surveys. *International Journal of Market Research 47*(6): 615–639.

Fahimi, M. 1994. *Post-stratification of pooled survey data. Proceedings of the American Statistical Association*. Survey Research Methods Section, Toronto, Canada.

Fahimi, M., D. Creel and P. Levy. 2007. *Evaluation of weighting methodology for the behavioral risk factor surveillance system. Joint Statistical Meetings (JSM).* Salt Lake City, UT.

Fahimi, M., D. Kulp and M. Brick. 2009a. A reassessment of list-assisted RDD methodology. *Public Opinion Quarterly 73*(4): 751–760.

Fahimi, M. and D. Kulp. 2009b. Address-based sampling – alternatives for surveys that require contacts with representative samples of households. *Quirk's Marketing Research Review,* May 2009.

Keeter, S., C. Kennedy, M. Dimock, J. Best and P. Craighill. 2006. Gauging the impact of growing non-response on estimates from a National RDD Telephone Survey. *Public Opinion Quarterly 70*(5): 759–779.

KnowledgePanel. 2015. Available at http://www.gfk.com/us/Solutions/ consumer-panels/Pages/GfK-KnowledgePanel.aspx.

Kott, P.S. 2006. Using calibration weighting to adjust for non-response and coverage errors. *Survey Methodology 32*(2): 133–142.

Kruskal, W. and F. Mosteller. 1980. Representative sampling IV: the history of the concept in statistics, 1895–1939. *International Statistical Review/Revue Internationale de Statistique 48*: 169–195.

Lee, S. 2006. Propensity score adjustment as a weighting scheme for volunteer panel web

surveys. *Journal of Official Statistics 22*(2): 329–349.

Neyman, J. 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society 97*(4): 558–625.

Rainie, L., S. Kiesler, R. Kang and M. Madden. 2013. *Anonymity, privacy, and security online.* Washington, DC: Pew Research Center's Internet & American Life Project. Available at: http://pewinternet.org/Reports/2013/Anonymity- online.aspx.

Skinner, C. 1999. Calibration weighting and non-sampling errors. *Research in Official Statistics 2*: 33–43.

Schonlau M., A. Van Soest and A. Kapteyn 2007. Are 'webographic' or attitudinal questions useful for adjusting estimates from web surveys using propensity scoring? *Survey Research Methods 1*(3): 155–163.

Smith, T.W., P. Marsden, M. Hout and J. Kim. 2013. *General social surveys*, 1972– 2012: cumulative codebook. National Opinion Research Center, Chicago.

Terhanian, G. and J. Bremer. 2012. A smarter way to select respondents for surveys? *International Journal of Market Research 54*(6): 751–780.