Quota Controls in Survey Research

A Test of Accuracy And Intersource Reliability in Online Samples

STEVEN H. GITTELMAN
Mktg, Inc.
steve@mktginc.com

RANDALL K. THOMAS
GfK Custom Research
Randall.K.Thomas@gmail.com

PAUL J. LAVRAKAS
Independent Consultant
pjlavrakas@centurylink.net

VICTOR LANGE
Catalina Marketing
vic.seriousemail@gmail.com

Editors' Note

In 2010, the Advertising Research Foundation (ARF) began a ground-breaking study—Foundations of Quality 2 (FoQ 2)—designed around several initiatives that investigated new questions about the quality of online samples. In addition, this research has examined how new sampling methodologies and technologies have evolved.

The latest FoQ 2 initiative in 2013 explored the effectiveness of quota controls in improving the quality of online survey data. The researchers investigated the use of traditional demographic quotas to balance samples to resemble probability (randomly selected) samples. The researchers also explored new model-based ways of selecting samples that used additional variables other than demographics. But the variability of results of one of the model categories (D) sparked differences among FoQ 2 committee members about how to interpret this variability, leading some to discourage including these results in this report.

In the end, the FoQ 2 committee agreed that publishing the work in its entirety was in the best interest of generating expanded research on this topic. Moreover, additional data will be needed to pursue a better understanding of which models will improve the accuracy of results using nonprobability samples, generally, and for specific topic areas.

Advertising Research Foundation (ARF) FoQ 2 Committee Leaders

Chris Bacon
Advertising Research Foundation
chris@thearf.org

Gian Fulgoni
comScore, Inc.
gfulgoni@comscore.com

George Terhanian
The NPD Group, Inc.
george.terhanian@npd.com

INTRODUCTION

Accuracy in survey research relies on accurate measurement of key topics with a sample that represents the population of interest. Can the way that an online sample is selected affect the accuracy of results?

Rather than conducting a survey on an entire population, almost all market research relies on selecting a subset of members of the population (a sample) to more efficiently represent the population. Whereas participants in a probability sample are randomly selected for participation at a calculable rate from a larger population, participants in a nonprobability sample (NPS) are not random representatives of the population of interest. Because of their significantly lower cost than probability samples, an NPS is most typical in the vast majority of online market-research studies. However, most nonprobability samples have been found to differ from the general population in specific ways—likely to be more highly educated, older, and with fewer minorities than probability-based samples (Chang and Krosnick, 2009).

A common method used to address the nonrepresentativeness of an NPS is to implement demographic quotas to make the sample demographically resemble the intended population. However, researchers have demonstrated that, even if demographics of an NPS are equated to the intended population, an NPS is neither behaviorally nor attitudinally the same (Baim, Galin, Frankel, Becker, and Agresti, 2009; Gittelman and Trimarchi, 2010b; Terhanian, Smith, Bremer, and Thomas, 2001; Yeager *et al.*, 2011).

In 2013, the Advertising Research Foundation's (ARF) Foundations of Quality 2 (FoQ 2) program implemented a large-scale study with 17 sample providers contributing nonprobability samples to an online survey that ran in parallel to a dual-frame phone survey. The main purpose of the investigation was to determine the extent of risk that existed with the use of nonprobability samples in online research.

The authors of the current paper examined two approaches to improve survey quality by making an NPS more representative of a general population:

DOI: 10.2501/JAR-2015-020

- implementing demographic sample-selection quota controls;
- using model-based selection methods, which used additional attitudinal and behavioral measures to select sample.

They measured quality in terms of the proximity of results to various U.S. benchmarks that have been established with large probability samples.

Generally, it was found that selecting samples with higher levels of demographic quota control did not increase data accuracy. In addition, and replicating results from FoQ 1 (Walker, Pettit, and Rubinson, 2009), nonprobability sample providers were not interchangeable with one another. Finally, although some model-based sample-selection approaches were found to reduce bias (improve proximity to benchmarks), approaching the accuracy of the telephone probability sample, there was no one method that provided universal improvement of the representativeness across all benchmarks that were used to assess accuracy.

The current article provides an in-depth account of the background, methodology, and results of this investigation.

BACKGROUND

In 2015, most online market research in the United States relied on nonprobability samples. Compared with a probability-based sample of the general U.S. population (*i.e.*, those drawn randomly to represent the U.S. population), participants in nonprobability samples have been found to be more active online, older, better educated, and less likely to have minorities (Chang and Krosnick, 2009; Dever, Rafferty, and Valliant, 2008).

Two primary methods have been used to make nonprobability samples demographically resemble probability samples drawn from the population of interest, with most methods focused on making the distributions of key demographic variables resemble the target population:

- **"Sample-Selection Balancing**" involves selecting sample members by using quotas that regulate the proportions of sample in various demographic groups. This process is followed so that the resulting sample has demographic distributions that approximate the distributions as they exist in the intended population (for such factors as age, sex, region of country, education level, or race/ethnicity groups). Sample-selection balancing requires that participants' demographics are either known prior to the survey or based on answers within a survey, most typically in an initial screening section.

- **"Demographic Weighting"** computes a multiplier for each respondent (each respondent is assigned a "weight") so that when the weights are applied, the multiplier will adjust for demographic proportions to match the distributions found in the target population:
  - ✧ Higher weight values increase the value of responses from participants from underrepresented groups;
  - ✧ lower weight values decrease the value of responses from participants from overrepresented groups.

Almost all online studies with nonprobability samples rely on one or both of these methods to balance samples demographically.

Studies that used sample-selection quotas typically have been found to yield similar demographics between nonprobability online research and more probabilistic methods (Gittelman and Trimachi, 2010a; Schonlau *et al.*, 2004; Terhanian *et al.*, 2001; Yeager *et al.*, 2011). Data from rigorously executed studies with nonprobability samples using quotas have been described as "well-behaved … with consequences far less pernicious" than many critics had stated (Stephenson, 1979).

Beyond Demographics

Even after making nonprobability samples resemble target populations demographically, nonprobability samples still have been found to significantly differ from probability samples, specifically in terms of attitudes and behaviors (Baim *et al.*, 2009; Baker *et al.*, 2013; Dever *et al.*, 2008; Piekarski *et al.*, 2008; Terhanian *et al.*, 2001; Yeager *et al.*, 2011). Compared with probability samples, demographically balanced nonprobability samples have been found to be:

- more active on the Internet
- more likely early adopters
- less traditional
- more environmentally concerned.

Without also taking into account and adjusting for these additional differentiating factors—especially when they are related to the area of study—the use of nonprobability samples could lead to different (and possibly misleading) results that could be costly for the researcher (Baker *et al.*, 2013). As a result of the non-demographic differences between probability and nonprobability samples, more complex model-based approaches have been developed to reduce the differences between nonprobability samples and target populations by taking into account the attitudinal and behavioral variables that differentiate the two (Baker *et al.*, 2013). These model-based approaches vary in terms of when they adjust the nonprobability sample to resemble the population:

- **"Model-Based Selection"** occurs when sample is selected (based on either prior or within-survey assessment of demographics, attitudes, and behaviors)

for a study to be similar to the target population proportions of all measured factors—both demographic and nondemographic (*e.g.*, Gittelman and Trimarchi, 2010a; Rivers, 2007; Terhanian and Bremer, 2012).

- **"Model-Based Adjustment"** creates weights or adjustment factors for the sample incorporating demographic, attitudinal, and behavioral information about the population (*e.g.*, Terhanian *et al.*, 2001; Terhanian, Bremer, and Haney, 2014).

Market-research studies vary in terms of how sample is selected according to demographic quotas. Some studies might have no quotas or they may impose only minimal quotas, controlling for the distributions of few key variables, such as age, sex, and region of country, but leave other variables uncontrolled (*e.g.*, race/ethnicity or education). Other studies might be more demanding in the quotas used for sample selection (*e.g.*, controlling not only for age, sex, and region but also for race/ethnicity and education distributions).

This article's authors expected that samples selected with fewer demographic quotas would produce less accurate population estimates (*i.e.*, higher bias with absolute differences from population values being greater) than those with more demographic quotas.

In comparison to demographic-based sample selection, however, the researchers expected that model-based methods that select samples for both demographic and nondemographic factors would render the nonprobability samples to be more similar to probability samples (*i.e.*, have greater proximity to national benchmarks).

METHODOLOGY

The main goals of the FoQ 2 sample-selection initiative were twofold:

- test the effectiveness of varying levels of sample selection;
- make recommendations on the sample-selection methodologies.

Online Survey

The FoQ 2 researchers requested that each of 17 online sample providers select three samples, each with more difficult sample-selection demographic quotas across an identical web-based omnibus questionnaire. Additionally, all 17 providers provided a single sample for one of the four model-based approaches that used both demographic and nondemographic variables to select sample (See Method D).

The sample-selection methods were as follows:

- "Method A" had sample-selection quotas for age and sex quotas nested within region based on U.S. Census values;

- "Method B" had Method A quotas plus race-ethnicity sample quotas based on U.S. Census values;

- "Method C" had Method B quotas plus education sample quotas based on U.S. Census values;

- For "Method D," there were four model-based approaches. Each provider involved in Method D sampling contributed a sample to only one of the four. Sample selection was based on:
  - ✧ demographic characteristics and
  - ✧ nondemographic attitudinal/behavioral variables, which included such factors as early adopter attitudes and community involvement.

Sample providers were provided with the numbers and proportions of completes within each strata for the demographic targets for each method. During fielding, providers were given daily feedback regarding the number and proportion of completes still required in each cell for each method.

The fielding of the general FoQ 2 online study took place between January 9, 2013 and January 24, 2013. The mean length of the online survey was 25.7 minutes. The number of completed online interviews (completes) was 70,377.

For the purposes of the current quota-control study, qualified completes averaged

- 1,072,
- 1,120, and
- 1,166

for Methods A, B, and C, respectively. The different numbers of completes by method reflected the somewhat increased difficulty of meeting the minimum number of participants with more stringent sample-selection quotas.

Sample Selection for Method D Model-Based Approaches

Thirteen of the providers contributed a sample to the one D model-based selection approach that screened sample within the survey for demographic and nondemographic factors. On the basis of this information respondents were selected to balance each provider's contribution to resemble a probability sample. Some providers contributed 500 completes, whereas others contributed 1,200 with an average of 740 completed interviews across all 13 providers.

For the other three D model-based approaches, respondents answered the key demographic and nondemographic questions specified in the models in a prior screening survey. Sample was invited to complete the survey after being selected from these pools of prescreened respondents for each model separately. Selection was done in such a way that would render

these nonprobability samples to be similar to probability samples with regard to demographic and nondemographic variables (*i.e.*, attitudes and behaviors). A single dedicated provider contributed 1,200 completes to two of these approaches, whereas two dedicated providers contributed about 500 completes to one of these approaches.

The authors described this sample-selection information without identifying the specific Method D approaches to maintain anonymity of the participating companies.

Parallel Telephone Survey

In addition to the nonprobability samples, the FoQ 2 study fielded a parallel telephone study with a subset of items used in the online study. (Fewer questions could be asked and answered in a telephone study of comparable length.) The phone study used a probability-based dual-frame methodology sample, with 40 percent of numbers coming from cell phones and 60 percent coming from landlines (completes, however, ended up as 31 percent cell/69 percent landline).

The parallel telephone study had 1,008 completed interviews during the field period of January 10–24, 2013, with a 19.9 percent response rate (AAPOR RR3). The mean length of the phone survey was 22.7 minutes. Both the online survey and phone survey were conducted in English only.

Poststratification Demographic Weighting

It was thought that sample selection (whether using quotas or model-based approaches) would help balance the nonprobability samples demographically. Yet poststratification weights were computed for each provider within each method separately (including the Method D models), with weighting factors of age within sex, region of country, race/ethnicity, and education. Additionally, the telephone survey data were weighted for these demographic factors along with the likelihood of contact by telephone.

RESULTS

Effects of Quotas on Weighting Efficiency

Precision of estimates is an important consideration when conducting research. A larger sample size is associated with a higher level of precision (lower margin of sampling error around the estimate). Weighting the results of a sample, however, increases variability of responses and lowers the effective sample size and thus lowers the precision of the results. In other words, when a sample is more similar to the population, weight variance is lower, leading to higher levels of weighting efficiency (and larger effective sample sizes). Higher weighting efficiency means that fewer respondents are required to obtain stable population estimates and are associated with greater statistical power overall.

The authors computed the weighting efficiency for each provider for each sample-selection method. Despite increasing levels of sample-selection quota control from Methods A to B to C, gains in weighting efficiency were inconsistent (See Figure 1):

- In the case of Method B, adding race-ethnicity quotas in addition to the age, sex, and region quotas did not improve weighting efficiency over Method A, which controlled only for age, sex, and region quotas while sampling.
- In the case of Method C, the addition of education sample-selection quotas along with age, sex, region, and race-ethnicity quotas significantly increased efficiency over Methods A and B.
- Two of the Method D model-based approaches (D-2 and D-4) also had significantly higher levels of weighting efficiency.
- Two other model-based approaches (D-1 and D-3) and the phone mode had lower levels of weighting efficiency.

Assessing Sample Bias—Distance from Benchmarks

The FoQ 2 survey had 29 questions (See Appendix) that paralleled benchmark variables from a variety of large-scale government and academic studies conducted with probability-based samples, including measures from the

- National Health Interview Survey
- American National Election Studies

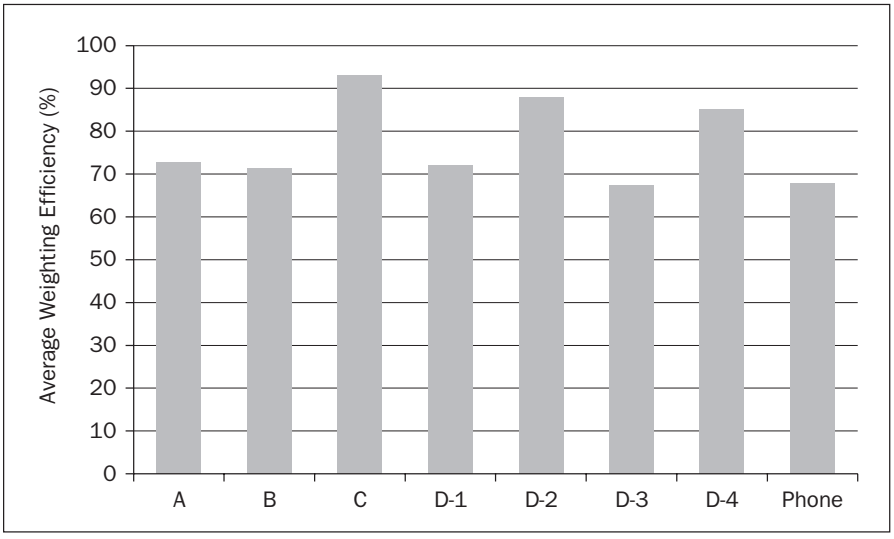

Figure 1 Average Weighting Efficiencies by Sample Method

- General Social Survey
- American Community Survey.

To ensure that the authors would not confound mode effects with sample effects in analyses, the authors examined the social desirability of the items. Social-desirability bias in responding is the tendency for survey respondents to either underreport undesirable activities or overreport desirable ones and is most likely to occur in situations with social presence—specifically results from telephone interviews are typically most affected (Baker *et al.*, 2013; Krumpal, 2013).

Five of the 29 questions were judged as higher in social-desirability bias, such as

- binge drinking frequency,
- mental health difficulties, and
- life satisfaction (See Gittelman *et al.*, 2015).

The authors excluded the five benchmark items subject to higher levels of social desirability from the analyses of this study. The authors noted that the specific five benchmark items had been assessed by a single national study using the telephone mode (the Behavioral Risk Factor Surveillance System [BRFSS]) and were not available with any other mode of assessment. This use of telephone interviewing made these particular national benchmarks most prone to be distorted by social-desirability bias and therefore would not serve as a clear indicator of sample quality.

Because the main purpose of this study was to examine differences due to sample rather than a mode difference, reporting for this study was based on the 24 items that were evaluated as being lower in social desirability.

All benchmarks used in the FoQ 2 study are summarized in the Appendix with the benchmark values that were obtained from all available national surveys (these surveys were administered in 2012 to early 2013, nearest to the field period for the FoQ 2 study). Some small differences between the sample estimates and national benchmarks were expected because the national benchmark studies had both English and Spanish language versions, whereas the online and phone surveys of FoQ 2 were in English only.

All benchmarks and estimates were proportions of those respondents who had specified substantive answers (*i.e.*, responses like "Decline to answer" or "Don't know" were not included in the computations). The authors' search for benchmarks yielded multiple sources for a number of variables, for which they took an average value across the national benchmarks for sample comparisons (See Appendix).

To test the accuracy of the FoQ 2 estimates, the researchers computed the absolute differences of the estimates obtained from the benchmarks for each of the 24 measures obtained from the online samples—the primary measure of sample bias. For example, if Provider 1 using Method A had an estimate of 17 percent for Alcohol Abstainer (Benchmark 1), and the benchmark value was 21 percent, the absolute difference of estimate was 4 percent. This was computed for each benchmark for each provider for each sampling method, and then averaged by provider within each method to derive an average absolute difference. The phone survey assessed 23 of the core 24 benchmarks.

Bias by Sample Method

The researchers first computed the average absolute differences between the national benchmark and the obtained item results by sample-selection method for both unweighted and weighted data (See Figure 2). Their expectations were that

- Unweighted data should show highest absolute differences from benchmarks with the fewest sample-selection quotas (Method A) and the highest absolute differences with the most sample-selection quotas (Method C).
- Beyond the sampling-selection quota methods, weighting may not have any effect if new cases of underrepresented groups (*e.g.*, minorities or lower education) have attitudes and behaviors just like already represented cases.

The research team, however, found that

- Increasing the extent of demographic sample-selection quotas did not generally lead to higher accuracy (no reduction of bias).
- Weighting the nonprobability samples did not reduce bias since it did not decrease average differences from benchmarks (See Figure 2).
- Among the Method D model-based sample-selection approaches, D-2 had the lowest average unweighted bias, the lowest bias overall, and a lower bias than the unweighted telephone survey.
- The unweighted results for the phone survey had the highest bias (highest average difference from benchmarks) across all samples, but weighting this probability sample significantly reduced bias (lower average differences from benchmarks).

Computing Margin of Sampling Error

An estimate based on a sample can be more precise when there are more participants, but it might not be accurate; in other words, the difference between the estimate and the population benchmark could have high bias. To examine the combined influence of precision (based on sample size) and bias (distance from benchmarks), the authors of the current paper next computed the margin of sampling error (using 95 percent confidence interval) for each item for each sample

provider within each method. Margin of sampling error[1] was chosen as the most appropriate method for comparing sample results with benchmarks since, by definition, it has a significance test associated with it based on sample result distribution and sample size.

The margin of sampling error for unweighted data used the actual number of completes, whereas the margin of sampling error for weighted data used the effective sample size based on the variance of the weights. The researchers used the margin of sampling error to determine whether the sample estimate was significantly different from the benchmark value for each item. This finding then was used to calculate the proportion of differences from benchmarks by provider and method that were significantly different.

The proportion of sample estimates (out of 24 for online, out of 23 for phone) was averaged within each method A, B, C, and D1-D4 (See Figure 3). Unweighted data had a smaller margin of sampling error because of larger sample size. Weighting reduces effective sample size, which generally reduces the number of comparisons that would be significant.

Using the 95 percent confidence interval for the computation of the margin of sampling error, the researchers had expected that about 5 percent of estimates on average would be significantly different from the benchmark solely due to sampling error. Besides sampling differences, there may be other reasons for a difference the authors did not test for, including

- measurement differences,
- context effects, or
- mode differences like social desirability.

[1] *"What is the Margin of Sampling Error?" Retrieved September 14, 2015, American Association for Public Research website: http://www.aapor.org/AAPORKentico/Education-Resources/For-Researchers/Poll-Survey-FAQ/What-is-the-Margin-of-Sampling-Error.aspx*

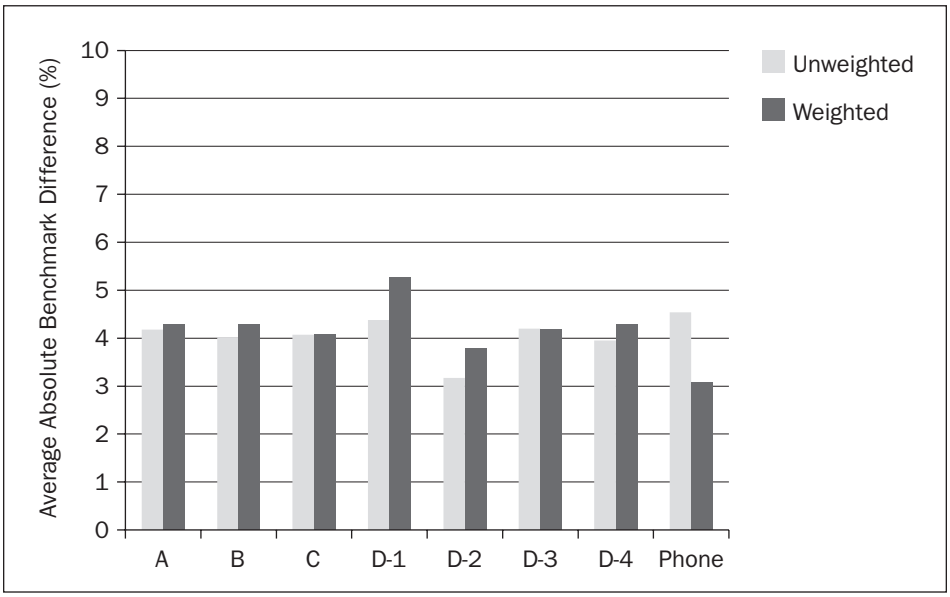

Figure 2 Bias by Sample Method—Average Absolute Differences from Benchmarks

Margin of Sampling Error Calculation

The researchers took the following steps for calculating margin of sampling error:

- They computed the proportion for each of the 24 benchmark questions for each method and provider (BP1 to BP24);
- They computed the margin of sampling error for the 95 percent confidence level for the unweighted data using the actual number of respondents; for example for BP1, the margin of sampling error would equal 1.96*SQRT((BP1*(1-BP1))/n);
- For the weighted data—since the sample size decreases due to weighting—they used the shrunken effective number of respondents resulting from weighting in the formula.
- The difference between the obtained unweighted/weighted proportions and the national benchmark was then computed by item within each method and provider—and compared to the margin of error to determine a "hit" (within the margin of error) or a statistically significant "miss" (outside the margin of error).
- The researchers then computed the proportion of significant "misses" (See Figures 3 and 6).

The authors found that:

- Each sample-selection method with unweighted results demonstrated deviations from population values well in excess of what the researchers had expected of a representative sample, with no method displaying less than 30 percent of significant differences from benchmarks (greater than the expected 5 percent). It should be noted, however, that the telephone survey—based on a probability sample—had one of the highest proportions of significant differences from benchmarks for unweighted data (above 60 percent).
- Among the demographic sample-selection methods, increasing demographic quota control was associated with higher levels of significant differences (from Methods A to B to C).
- In all but one case, weighting decreased the proportion of items showing

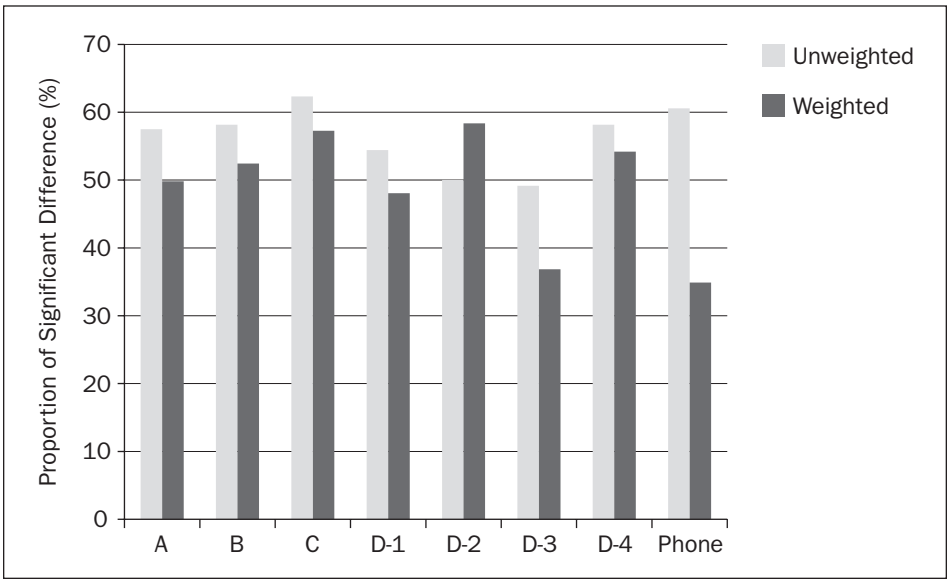

Figure 3 Average Proportion of Significant Differences From Benchmarks by Method

significant differences, meaning that fewer benchmarks fell outside their respective confidence intervals.

- Unlike the results found for average absolute differences (See Figure 2: no real differences in bias), however, the proportion of significant differences increased going from Methods A to B to C.

Some reasons underlying the results that the authors found for the sample-selection methods may be a consequence of both the sample size required to meet the minimum sample-selection quotas and the weighting efficiency for each method. Specifically,

- Method B required an average of 48 (about 4.4 percent) more respondents per provider than Method A due to increased difficulty filling quotas, and
- Method C required 46 (about 4 percent) more respondents than Method B.

With increasing sample size, the margin of sampling error became smaller (indicating greater precision), increasing the frequency with which smaller differences between sample estimates and population benchmarks were found significant. Therefore, the authors of the current article believe that the proportion of significant differences that increased from A to B to C may have occurred partially as a result of an artifact of the increased sample sizes.

Further, the reduction of the proportion of significant differences, due to weighting by method, may be partly due to a reduction of the effective sample size that occurs with weighting. A reduction of effective sample size would increase the margin of sampling error (indicating less precision), which, as a result, made fewer comparisons with benchmarks significantly different.

Method D Findings

In looking at the Method D model-based sample-selection approaches:

- Unweighted data from Models D-1, D-2, and D-3, had a lower average proportion of significant differences when compared with D-4, the phone method, and Methods A, B, and C.
- After demographic weighting, however, only D-3 and the phone method had the lowest and most comparable proportion of significant differences from national benchmarks.

Method D-3, however, also had the fewest average completes (about 740), so part of the effect may have been an artifact of the way significant differences using margin of sampling error were calculated. Fewer completes are associated with a wider margin of error (lower average precision), and, therefore, fewer differences as a proportion of all differences will be significant (or outside the margin of sampling error).

Findings by Benchmark for Nonprobability Samples

The researchers next examined each of the specific measures of the national benchmarks to determine the average extent of difference of sample estimates from the benchmarks for results from Methods A, B, and C only (See Figure 4).

- Four items, in particular, had an average of 6 percent difference or greater between the sample estimates and benchmarks (See Appendix: Items 4, 10, 16, and 17).

- Specifically, nonprobability samples indicated
  - ✧ a higher proportion of current smokers,
  - ✧ lower rates of strong religious strength,
  - ✧ lower rates of full-time employment, and
  - ✧ lower rates of home ownership compared with their respective benchmarks.

These particular measures of benchmarks may reflect real sample-level differences between nonprobability samples and the general population and do not appear to be eliminated by demographic weighting. With differences of this magnitude, however, the researchers could not rule out alternative influences—including:

- measurement differences,
- some types of mode effects, or
- social desirability (*e.g.*, See Appendix, Item 10: [Religious Strength was higher in the benchmark than for online probability samples]).

The authors also noted that the four specific items with the highest unweighted bias (biggest difference from benchmarks) did not have a reduction of bias when weighted and, in fact, for three items, weighting increased bias.

Findings by Benchmark for Sample-Selection Method

Next, the researchers examined the extent of bias (absolute average deviation from benchmarks) by each sample-selection method (See Table 1):

- Methods A, B, and C did not show big differences.
- Method D model-based approaches varied in their performance over a very wide range of benchmarks.
- Some of the smallest differences between sample estimates and benchmarks were not found with Methods A, B, or C but instead were found with some of the Method D approaches.

This implies that Method D approaches may have differential effectiveness, depending on topic.

Both Methods D-2 and D-4, in fact, had eight comparisons (out of the core 24 benchmark items) that were lowest of all the methods tested, providing above-average performance. There currently does not appear, however, to be a "one-size-fits-all" solution for model-based sample selection (and, by implication, model-based sample adjustment). Different methods use different combinations of covariates generating the patchwork-quilt performance noted here (See Table 1).

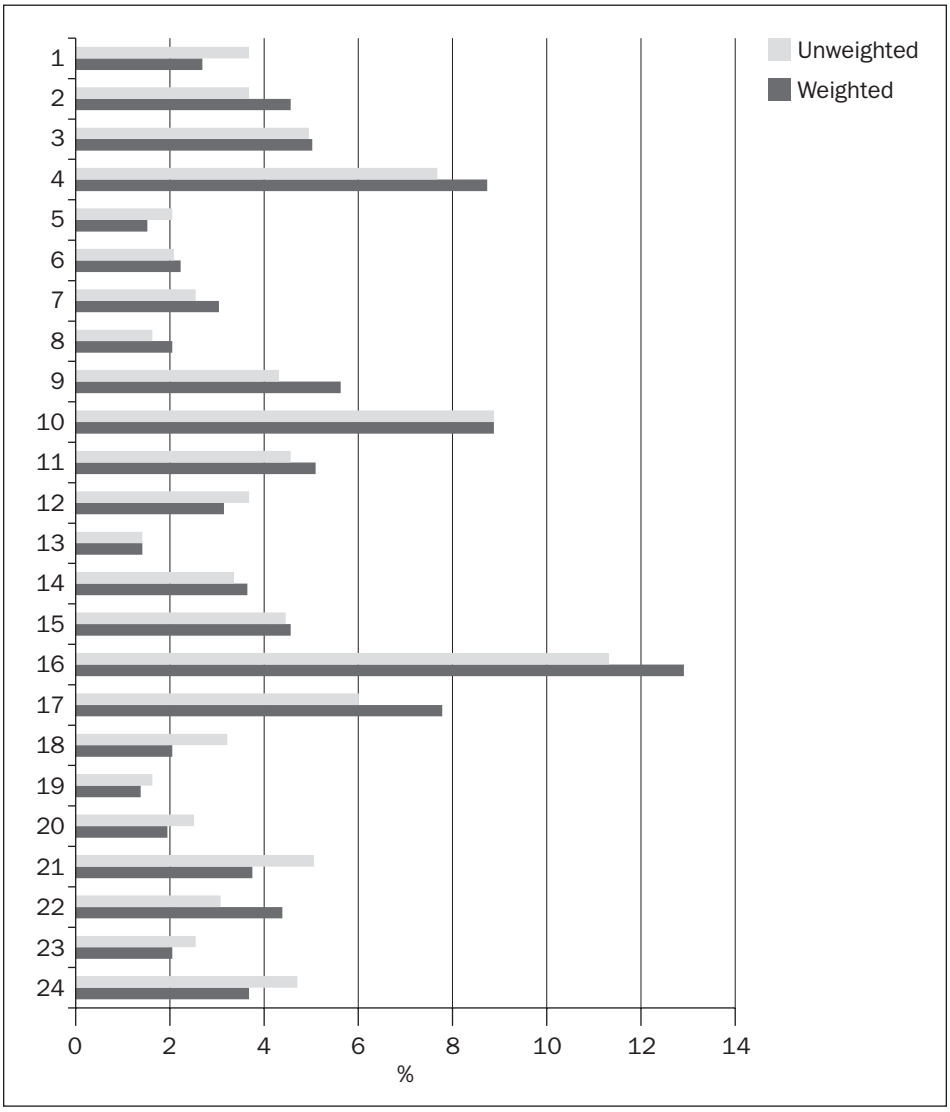

Figure 4 Bias Across Providers
(Average Absolute Deviation for Each Benchmark)

Sample Provider Differences

The researchers next examined differences between the 17 providers in their bias (average absolute deviation from benchmarks) for samples selected with Methods A, B, and C only. The phone method sample results, a probability sample, also were presented to enable comparison with the other providers (See Figure 5). Among the findings:

- The weighted telephone sample had an average 3 percent difference from benchmarks.
- Eight of the 17 providers had between 3 percent and 4 percent weighted average deviations from the benchmarks.
- Five of the providers had weighted average deviations of 5 percent or greater from the benchmarks.
- Demographic weighting reduced bias for only five of the 17 nonprobability sample providers, whereas weighting increased bias for eight of the 17 samples.

These results indicated that although some providers appeared to have provided

TABLE 1

Bias for Benchmarks by Sample Method (Average Absolute Difference From Benchmark)

Benchmark Number	Benchmark	A (%)	B (%)	C (%)	D-1 (%)	D-2 (%)	D-3 (%)	D-4 (%)	Phone (%)
1	Alcohol Abstainer	2.7	2.4	2.8	6.0	2.5	3.0	2.1	0.3
2	Current Regular Drinker	4.2	5.1	4.4	6.6	8.5	5.2	4.7	7.9
3	Never Smoked Cigarettes	5.5	4.9	4.5	7.1	4.2	5.0	5.3	0.5
4	Current Smoker	10.3	8.0	7.8	8.5	5.4	8.2	13.1	4.6
5	Below Average Sleep	1.4	1.7	1.4	3.5	0.4	3.0	1.2	2.2
6	Is Obese	2.5	2.3	1.7	3.7	1.7	3.2	0.4	
7	Good Self-rated Health	3.0	3.1	3.0	4.2	5.4	3.1	6.6	0.9
8	Conservative Ideology	1.4	2.8	1.8	4.1	2.2	2.3	1.6	0.2
9	Republican Party ID	5.3	6.1	5.5	5.3	1.0	5.1	4.7	4.8
10	Strong Religious Strength	8.8	9.0	8.7	9.1	8.9	3.7	6.4	2.7
11	High Religious Attendance	5.1	5.3	4.8	5.3	3.2	2.5	0.8	4.4
12	Landline Phone in Household	3.9	2.7	2.7	0.6	2.2	4.5	3.3	0.6
13	Cell Phone in Household	1.0	1.6	1.4	3.0	0.2	1.6	0.4	1.2
14	Married	3.6	3.8	3.4	5.5	3.5	2.7	6.2	1.0
15	Not Employed	4.0	4.9	4.6	2.8	3.0	4.6	7.2	1.7
16	Full-time Employed	13.7	12.6	12.4	14.9	12.3	14.2	18.3	8.0
17	Own Home	7.7	7.6	7.9	10.1	10.2	8.0	9.9	4.5
18	Speak Only English at Home	2.1	1.8	2.1	1.5	3.0	2.7	0.2	0.4
19	Vehicle in Household	1.6	1.1	1.3	2.0	1.6	1.8	0.0	1.7
20	4+ Bedrooms in Residence	2.0	2.0	1.7	2.9	1.3	2.2	0.9	10.1
21	Children in Household	3.4	4.1	3.6	8.0	4.7	3.6	1.2	2.6
22	Registered to Vote	3.9	4.7	4.5	6.1	2.2	4.3	3.8	5.5
23	Has Valid Driver's License	2.2	2.0	1.7	1.5	1.3	2.7	0.7	2.2
24	Has Valid Passport	3.3	3.4	4.2	4.4	1.8	3.3	4.1	2.7

nonprobability samples that were much more similar to what would be obtained from a probability sample, others were more significantly different.

The research team also compared providers (for Methods A, B, and C only) and the telephone sample in terms of the average proportion of significant differences using the calculated margin of sampling error across benchmarks (See Figure 6):

- Except for one nonprobability sample (Provider 3), all samples had unweighted proportions of significant differences over 50 percent.
- The weighted telephone sample had 34.8 percent of comparisons significantly different from benchmarks using the margin of sampling error calculations.
- One provider (Provider 11) was lower than the telephone data with weighted data with 31.9 percent of results significantly different from benchmark values.

CONCLUSION AND DISCUSSION

Regarding sample-selection methods, as a result of this study:

- Method A selection may have been an adequate selection strategy for nonprobability sample. Increasing the extent of demographic selection quotas used did not reduce bias or improve accuracy (Method C was no more accurate than Method B which was no more accurate than Method A).
  - ✧ Adding race-ethnicity and education quotas did not reduce bias (proximity to the benchmarks was not reduced).
  - ✧ The costs of increased use of more complex demographic selection quotas may not be justified in terms of reducing bias of responses for nonprobability samples, at least across the wide range of benchmarks used in this study.
  - ✧ The primary utility of oversampling underrepresented groups enables results to be obtained for these groups, but oversampling did not appear to decrease bias.

- Demographic weighting did not generally reduce bias for nonprobability samples when using demographic sample selection.

- Some Method D model-based sample-selection approaches showed promise in reducing bias and improving accuracy, as previously suggested (Baker *et al.*, 2013). The current study showed that variables other than demographics could be useful in reducing the differences in results between nonprobability samples and probability-based estimates.

- Although this study focused on a general population sample, a disadvantage of Method D sampling approaches was that they generally required considerably more sample to screen than was selected for participation (whether it occurred within the survey screening section or as part of preprofiled sample).
  - ✧ The one Method D sample-selection approach that used in-survey respondent screening to select a balanced sample for each provider required about 40 percent more sample on average than Method A.
  - ✧ The other Method D approaches also may have required an equivalent number of participants to be prescreened in order to select a balanced sample.

- As was found in the FoQ 1 study (Walker *et al.*, 2009), nonprobability sample providers were not interchangeable: Some appeared to provide samples that were close to probability samples across a range of metrics, whereas others did not.

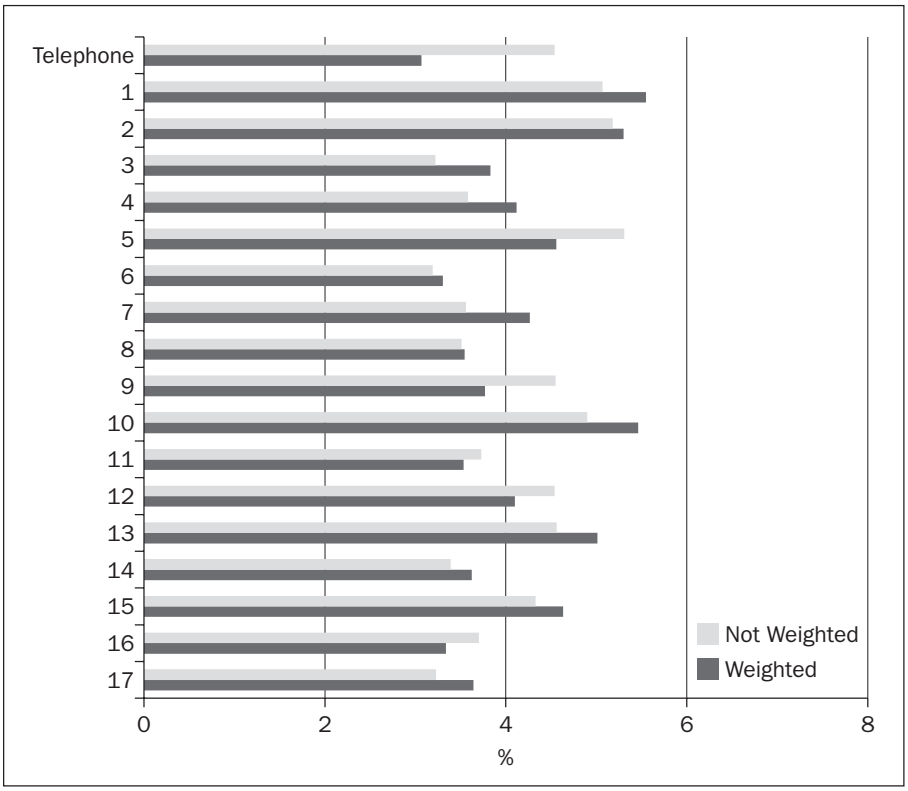

Figure 5 Average Absolute Benchmark Difference for Telephone and by Online Provider

Future Directions

Model-based approaches—to sample selection or post-fielding adjustment—that use attitudinal or behavioral variables, in addition to demographic variables, hold promise and should be further explored. No matter how exacting the demographic quotas are for a given study, the current authors believe, traces of the bias inherent in nonprobabilistic samples often will survive the demographic quota process, depending on the topic being investigated.

The goal of future research would be to understand

- both the general and subject-specific selection or adjustment variables;
- how the nature and extent of the correlation of the adjustment/selection variables affects results with substantive items;
- the limits and range of applicability of the models to the topics of interest.

The current authors end with a note of caution to practitioners. The Method D model-based selection approaches depend on using variables in their selection (or adjustment) algorithms that co-vary with study variables, or they likely will not work. The question of "how close is too close" in terms of the selection variable and the substantive variables still must be studied in the future to better understand when and how it is best to use them.

Reinforcing the potential dangers that threaten practitioners, the authors cited a metaphor from Greek mythology:

> *Daedalus, a skillful craftsman and inventor, and his son, Icarus, were prisoners of King Minos on the Island of Crete. Daedalus created wings of wax and feathers so that he and his son could fly to freedom. Despite his father's cautions, Icarus flew too close to the sun. The wax melted and, to Daedalus' horror, Icarus fell to his death.*

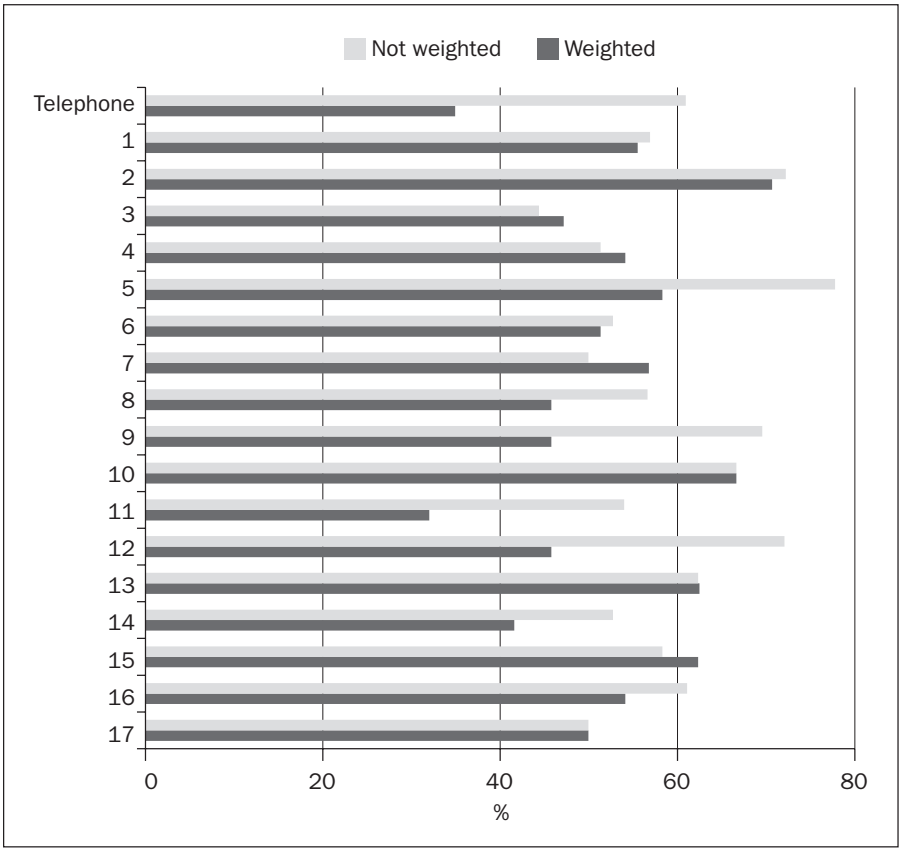

Figure 6 Average Proportion of Significant Differences From Benchmarks by Provider

Fortunately, in regard to the current study, the consequences are not fatal. But those practitioners who do use the D methods in the studies they perform risk falling victim to high correlations between the selection variable and substantive variables. Future research should address the question of how best to balance these concerns. JAR

ABOUT THE AUTHORS

Steven H. Gittelman is president and chief methodologist of Mktg, Inc., a marketing-research firm in East Islip, NY. His research specialty is online-survey quality. Gittelman has published numerous articles and four books, including *J.P. Morgan and the Transportation Kings: The Titanic and Other Disasters* (University Press of America, 2012).

Randall K. Thomas is vice president of online research methods at GfK Custom Research. With more than 25 years of experience in conducting survey projects across multiple modes and across countries—including research positions at Harris Interactive and ICF International—Thomas has pursued multiple lines of research into measurement accuracy of attitudes, intentions, and behaviors in web-based surveys. His work can be found in more than 20 publications and 170 conference presentations.

Paul J. Lavrakas is a research psychologist, a research methodologist, and an independent consultant. He also is a senior fellow at NORC, the University of Chicago's social-science research organization, and a visiting scholar at Northern Arizona University. Lavrakas is the editor of the *Encyclopedia of Survey Research Methods* (Sage, 2008), co-author of *Applied Qualitative Research Design* (Guilford Press, 2015), as well as the author of many other books, chapters, and articles on various aspects of research methodology.

Victor Lange is a research analyst at Catalina Marketing in Saddle Brook, NJ. Previously, and while this article was being written, he was a statistical analyst at Mktg., Inc., focusing on devising and evaluating best practices in online survey research.

REFERENCES

Baim, J., M. Galin, M. R. Frankel, R. Becker, and J. Agresti. "Sample Surveys Based on Internet Panels: 8 Years of Learning." Paper presented at the Worldwide Readership Symposium, Valencia, Spain, October 2009.

Baker, R., J. M. Brick, N. A. Bates, M. Battaglia, *et al.* "Summary Report of the AAPOR Task Force on Non-Probability Sampling." *Journal of Survey Statistics and Methodology* 1, 2 (2013): 90–143.

Chang, L., and J. A. Krosnick. "National Surveys Via RDD Telephone Interviewing Versus the Internet: Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 73, 4 (2009): 641–678.

Dever, J., A. Rafferty, and R. Valliant. "Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?" *Survey Research Methods* 2, 2 (2008): 47–60.

Gittelman, S., and E. Trimarchi. "Online Research… And All That Jazz! The practical adaptation of old tunes to make new music." Paper presented at ESOMAR Online Research Conference, Berlin, Germany, October 2010a.

Gittelman, S., and E. Trimarchi. "The Perfect Storm: Strong Impacts on Buying Behaviour." *Vue Magazine*, May 2010b.

Gittelman, S. H., V. Lange, W. A. Cook, S. M. Frede, *et al.* "Accounting for Social Desirability Bias: A Model for Predicting and Calibrating the Direction and Magnitude of Social Desirability Bias." *Journal of Advertising Research* 55, 3 (2015): 242–254.

Krumpal, I. "Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review." *Quality & Quantity: International Journal of Methodology* 47, 4 (2013): 2025–2047.

Piekarski, L., M. Galin, J. Baim, M. Frankel, *et al.* "Internet Access Panels and Public Opinion and Attitude Estimates." Presented at 63rd

Annual AAPOR Conference, New Orleans, LA, May 2008.

Rivers, D. "Sampling for Web Surveys." Paper presented at the Joint Statistical Meetings, Salt Lake City, UT, August 1, 2007.

Schonlau, M., K. Zapert, L. Payne Simon, K. Sanstad, *et al.* "A Comparison Between Responses From a Propensity-Weighted Web Survey and an Identical RDD Survey." *Social Science Computer Review* 22, 1 (2004): 128–138.

Stephenson, C. B. "Probability Sampling With Quotas: An Experiment." *Public Opinion Quarterly* 43, 4 (1979): 477–496.

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

Terhanian, G., J. Bremer, and C. Haney. "A Model Based Approach for Achieving a Representative Sample." Paper presented at CASRO Digital Research Conference, San Antonio, TX, March 2014.

Terhanian, G., R. Smith, J. Bremer, and R. K. Thomas. "Exploiting Analytical Advances: Minimizing the Biases Associated With Non-Random Samples of Internet Users." Proceedings from ESOMAR/ARF Worldwide Measurement Conference (2001): 247–272.

Walker, R., R. Pettit, and J. Rubinson. "A Special Report from the Advertising Research Foundation: The Foundations of Quality Initiative: A Five-Part Immersion Into the Quality of Online Research." *Journal of Advertising Research* 49, 4 (2009): 464–485.

Yeager, D. S., J. A. Krosnick, L. Chang, H. S. Javitz, *et al.* "Comparing the Accuracy Of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples." *Public Opinion Quarterly* 75, 4 (2011): 709–747.

APPENDIX

Benchmark Number	Benchmark	Criterion	Benchmark Value (%)
1	Alcohol Abstainer	Never drank 12 or more drinks in total	21.0
2	Current Regular Drinker	Drank at least 12 drinks in past year	51.6
3	Never Smoked Cigarettes	Never smoked 100 cigarettes or more	55.9
4	Current Smoker	Smoke some days or every day	18.5
5	Below Average Sleep	Less than 6.5 hours/night	32.0
6	Is Obese	BMI greater than or equal to 30	31.3
7	Good Self-rated Health	Good or better	82.6
8	Conservative Ideology	Lean Conservative or more	41.0
9	Republican Party ID	Lean Republican or more	37.5
10	Strong Religious Strength	Very or Moderately Religious	58.3
11	High Religious Attendance	Almost every week or every week	32.5
12	Landline Phone in Household	Has at least 1 in household	65.0
13	Cell Phone in Household	Has at least 1 in household	91.9
14	Married	"Yes" to married	52.9
15	Not Employed	"No" to all employment questions	40.6
16	Full-time Employed	Equal to or greater than 35 hours worked	45.9
17	Own Home	Own (with or without mortgage)	67.1
18	Speak Only English at Home	"No" to speaking language other than English	79.2
19	Vehicle in Household	One or more vehicles	90.8
20	4+ Bedrooms in Residence	Four or more bedrooms	20.2
21	Children in Household	One or more children under 18	34.5
22	Registered to Vote	Indicated registered for current or prior place	87.1
23	Has Valid Driver's License	Indicated has valid driver's license	86.2
24	Has Valid Passport	Indicated has valid passport	40.6