Marketing Systems Group

CALIBRATING NON-PROBABILITY INTERNET SAMPLES WITH PROBABILITY SAMPLES USING EARLY ADOPTER CHARACTERISTICS

DiSogra C., C. Cobb, M. Dennis, 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*

Abstract

A representative study sample drawn from a probability-based Web panel, after post-stratification weighting, will reliably generalize to the population of interest. Due to finite panel size, however, there are instances of too few panel members to meet sample size requirements. In such situations, a supplemental sample from a non-probability opt-in Internet panel may be added. When both samples are profiled with questions on early adopter (EA) behavior, opt-in samples tend to proportionally have more EA characteristics compared to probability samples. Taking advantage of these EA differences, this paper describes a statistical technique for calibrating opt-in cases blended with probability-based cases. Using data from attitudinal variables in a probability-based sample (n=611) and an opt-in sample (n=750), a reduction in the average mean squared error from 3.8 to 1.8 can be achieved with calibration. The average estimated bias is also reduced from 2.056 to 0.064. This approach is a viable methodology for combining probability and non-probability Internet panel samples. It is also a relatively efficient procedure that serves projects with rapid data turnaround requirements.

Key Words: Calibration, Web surveys, online panels, probability-based samples, opt-in samples, Internet panels

Introduction

Internet surveys are becoming an increasingly popular mode of data collection in public opinion research. Technological progress and the rising penetration of the Internet in everyday life means a large group of people can be reached quickly and from whom answers can be rapidly collected and analyzed. Internet surveys are generally less expensive to administer than telephone and in-person surveys (Fricker & Schonlau 2002) and Internet surveys using probability-based samples have been shown to yield estimates as accurate as or more accurate than other survey modes. Probability-based Internet panels use a traditional survey sampling frame, such as, random-digit dial (RDD) or an address-based sample frame (ABS). All households in the frame have a known, non-zero probability for selection. KnowledgePanel®, the Knowledge Networks (KN) online panel, initially used RDD but now employs ABS. For households that have no Internet access, KN provides a laptop computer

and Internet service to allow their participation. This achieves more complete population coverage on this nationally representative panel for conducting survey research projects.

However, due to finite size, nationally representative probability-based Internet panels sometimes have too few panel members to meet sample size requirements for studies interested in small geographic areas, special sub-populations or rare incidence phenomena. When such conditions arise, non-probability opt-in Internet panel cases may supplement the available probability samples to obtain an overall sample size large enough to study the topic or group of interest. Because opt-in respondents may be less representative than probability-based respondents, it is necessary to correct for bias from the opt-in sample component when combining data.

Calibration weighting is a class of techniques for combining data from different sources and is often used to correct for bias (Kott 2006; Skinner 1999). However, it is typically overlooked as a cumbersome, multi-step process that can be costly in terms of time and money for a researcher to utilize. The primary purpose of this paper is to describe a more efficient calibration weighting approach for use when blending Internet survey data from probability and non-probability samples that is cost effective and useable across a number of different study topics and sample populations. Moreover, we have empirically evaluated its effectiveness on study estimates by making comparisons to other approaches for combining these two types of samples.

For background, Section 2 will briefly describe differences in data quality between probability and non-probability Internet samples and points to the need for calibration when combining data from both sources. Section 3 contains a brief literature review of calibration. Section 4 identifies a series of five questions related to attitudes toward the early adoption of new technology and products that we have found to consistently differentiate probability-based respondents from opt-in respondents across demographic groups. It is with these early adopter characteristics that we adjust or “calibrate” the opt- in sample to minimize bias. Section 5 gives step-by-step instructions on how KN performs calibration. And lastly, Section 6 contains the results of an evaluation of our calibration technique by examining its impact on the mean squared error of 13 attitudinal survey questions.

Probability Recruited Samples and Non-Probability Opt-in Samples

Two types of Internet panels exist by which to estimate the opinions and behaviors of the general public. One uses a probability-based recruitment approach employing either RDD or ABS frames. These sampling frames provide nearly all households with a known non- zero chance of being included on the panel. Recruited households without Internet access can be provided the necessary equipment, access, and support to participate in online surveys (this is the KN model). Completion rates are usually high (averaging between 65 to 70%). Results are generalizable and can be used to calculate prevalence estimates with applicable confidence intervals. Probability-based Internet panels are currently used extensively by government,

academic, and industry researchers in studies where a high degree of rigor is desired. These types of panels are recognized by the American Association of Public Opinion Research (AAPOR) as a valid and reliable survey method (AAPOR 2010). Moreover, a number of studies have found results from probability- based Internet panels to have higher concurrent validity, less survey satisfying, and less social desirability bias than telephone surveys (Couper 2000; Chang & Krosnick 2009; Kypri, Stephenson & Langley 2004).

The second type of Internet panel is a non-probability opt-in panel whereby respondents are recruited through Internet advertisements, recruitment Websites or email invitations based on commercial lists. Persons are not "selected" to be recruited; it is solely on their proactive interest in joining such panels (usually for monetary compensation) that they exercise their option to be a member. Ergo, these types of panels are commonly called "opt-in" panels. Opt-in panels are frequently used by market researchers because of their relatively low cost and greater ability to target defined types of respondents due to very large membership numbers (often in the millions). However, the members of these opt-in Internet panels have no known probability of selection as they self-select from a pool that can only be described as "persons on the Internet." Such panels are limited further because the population without Internet access is excluded. While exact recruitment, sampling, and weighting methods for commercial opt-in panels are often not transparent and treated as proprietary information, attempts to overcome potential sample bias likely include quota sampling of various degrees of complexity and/or extensive post-survey adjustments, albeit with questionable success. Fundamentally, they are convenience samples.

Yeager et al. (in press) compared survey estimates from seven opt-in Internet panels, a probability-based Internet panel, and an RDD (probability) sample. While the RDD sample and the probability-based Internet panel were "consistently highly accurate" with average absolute errors of only 2.9% and 3.4%, respectively, the opt-in panels were always less accurate with average absolute errors ranging from 4.5% to 6.6%. Post- survey weighting adjustments even worsened the accuracy of one opt-in sample. Furthermore, Yeager et al. found no association between accuracy on any one benchmark to the overall accuracy on all of the benchmarks for opt-in samples. While Yeager et al. is the most extensive study to date to raise serious concerns about estimates produced by opt-in panels, it is not alone (see Bethlehem & Stoop 2007; Couper 2000; Lee 2006; Pasek & Krosnick 2010). AAPOR even considers it "harmful" and "misleading" to report a margin of sampling error for estimates produced from opt-in panel surveys (AAPOR 2011).

Probability-based Internet panels are capable of providing data of high quality that are generalizable to the larger population. However, because of recruitment costs, the current sizes of national probability-based panels, usually in the tens of thousands, can be a barrier for using them with projects that are interested in very small geographic areas or rare populations/behaviors. Opt-in Internet panels, on the other hand, with their millions of panelists, have sufficient panel sizes to study smaller geographic areas and rare populations, but yield estimates of lower quality and greater bias. The result is that both probability-based Internet panels and opt-in Internet panels individually may not be sufficient for some studies.

However, blending the two together using a calibration technique can take the relative advantages of each to produce estimates that are much closer in precision to the estimates one would expect if a larger probability-based panel was available. In the next section, we discuss a calibration solution that is KN's approach to the technique.

Calibration Weighting

Calibration weighting has been part of the survey researcher's toolkit for a long time (Kott 2006; Sarndal 2007; Skinner 1999). It is a collection of techniques that attempt to correct for coverage bias in survey samples by adjusting sampling weights by multipliers that make estimates agree with known population totals. It is at its essence an extension of the well-known practice of coverage adjustment through post-stratification. The basic idea is to take estimates from one source of data, which may themselves be sufficiently accurate population estimates, to use as "benchmarks" to adjust the estimates of the less accurate source of data. The result is a larger data set with its corresponding advantages for analytical purposes.

Calibration weighting techniques share a number features that make them useful for combining data from multiple surveys. Calibration provides a systematic way to use auxiliary information that is different between the two samples to improve the accuracy of survey estimates (Reuda et al. 2007). Auxiliary data can come from multiple surveys and can exist at either the aggregate or individual levels (Sarndal 2007). Attitudinal or lifestyles questions, which are usually unavailable from the census demographic data that are most often used to calculate survey weights, can capture the difference between opt-in respondents and probability-based respondents, even when they are unrelated to the survey topic. Calibration weighting then provides a method for correcting estimates based upon those differences.

Another advantage of calibration weighting compared to other methods addressing coverage error reduction is that calibration weighting invokes no assumptions about data or modeling. Calibration on known totals is easy to understand for most researchers familiar with sample weighting. It proceeds by slightly modifying post-stratification weights to reproduce the specified auxiliary totals. There is no need to explicitly state a model of the relationship between the auxiliary variables and the probability of being included in each of the samples. Other methods of correction that rely on an explicit model, such as propensity score adjustment, can actually introduce bias into the blended data if the model underlying the adjustments is mis-specified (Guo & Fraser 2010). Furthermore, propensity score adjustments require that the likelihood of being in the opt- in sample be independent of any outcomes of interest, which one does not know empirically until after data collection is complete and can require different models for different variables of interest in the same survey (Rosenbaum & Rubin 1986; Schonlau, van Soest & Kapteyn 2007). Calibration weighting, on the other hand, can be done using any variable that differentiates between the two samples irrespective of its relationship to any other variable, does not require the considerable amount of analyst time and effort to specify an optimal model, and provides a single adjustment solution that is

useable to produce less biased estimates for all the variables of interest in the data.

The calibration weighting approach to blend a probability-based sample with an opt-in sample demonstrated in this paper is different from other calibration approaches in two important ways. First, the survey administered to the probability-based sample, from which calibration benchmarks are taken, is identical in mode and design to the survey administered to the opt-in sample. The mode via which a survey is administered is known to affect the data generated from it (Dillman 2000; Goyder 1985; de Leeuw & van der Zouwen 1988). Small differences in the wording of survey questions are known to lead to very different distributions of answers. By administering the same set of questions to all respondents in both the probability sample and non-probability sample in same mode, the real difference between the samples on benchmark variables is more accurately measured (rather than artificially over- or understated due to mode differences), which in turn improves the accuracy of weight adjustments obtained from the calibration process.

Second, the approach outlined below further reduces analyst burden and the time it takes to calibrate by eliminating unnecessary steps related to post-stratification weighting of the opt-in sample data prior to calibration. Calibration is itself an extension of post- stratification adjustment and the benchmarks that differentiate the probability and non- probability samples become part of the statistical raking procedure, along with standard demographics, that creates the final weights (Sarndal 2007). As we will demonstrate later, since the overall goal is to make the blended sample resemble known population totals, there is no improvement to first weight the non-probability sample to the same known population totals. The same logic does not apply to pre-weighting the probability- based sample, as will be discuss below, since it is also the source of the population benchmark estimates used for calibration. It is worthwhile to note that when designing the opt-in sample it is ideal to use some demographic quotas, if feasible, to minimize the size of the inevitable weighting adjustments.

Calibration Benchmarks: Early Adopter Characteristics

The auxiliary variables needed for calibrated weighting must reliably differentiate between the probability-based sample and the opt-in sample. Dennis et al. (2009) looked for differences in attitudinal questions between four U.S.-based general population Internet panels and found that opt-in panels demonstrated higher proportions of respondents who are likely to report attitudes aligned with being early adopters of new products and concepts.

Early adopters (EA) are defined as consumers who embrace new technology and products sooner than most others. It is a consumer segment that has been used by marketing research since the 1950s when they were first identified in Francis S. Bourne's seminal essay "The Adoption Process" (Bourne 1959). Marketers are particularly interested in this group because EA consumers are willing to spend money at an early stage and can encourage the spread of new products among their friends and colleagues. However, as Dennis et al. (2009, p.2)

rightly warn "If a survey sample consists of too many *early adopters*, the survey might provide inflated and erroneous measures of willingness to purchase…, leading to bad business decisions."

The Dennis et al. study fielded identical surveys to respondents from the following four Internet panels: the 2007-2009 American National Election Studies (ANES) Web Panel, a probability-based recruited panel whose main purpose was academic research and funded by a grant from the National Science Foundation; KN's KnowledgePanel, a privately owned probability-based recruited panel used for commercial, government and academic research; and Opt-in Web Panels A and B, both of which are randomly selected from a list of well-known opt-in panel firms.

Table 1, reproduced from Dennis et al., presents differences in the proportion of respondents from each panel who agree/strongly agree with each of the five EA statements (EA1-EA5).

Table 1: Percent of respondents agreeing/strongly agreeing with early adopter (EA) statements by panel source.

	ANES Web Panel	Knowledge Panel	Opt-in Web Panel A	Opt-in Web Panel B
I usually try new products before other people do (EA1)	26.4	24.0	44.2*	41.4*
I often try new brands because I like variety and get bored with the same old thing (EA2)	36.6	34.1	52.0*	54.2*
When I shop I look for what is new (EA3)	44.5	35.7*	55.2*	59.0*
I like to be the first among my friends and family to try something new (EA4)	23.8	22.2	38.1*	39.6*
I like to tell others about new brands or technology (EA5)	51.8	45.0*	60.2*	62.1*
Sample size	1,397	1,210	1,221	1,223
Completion Rate	65.8%	63.7%	4.6%	4.7%

Difference compared to ANES Web Panel uses Fisher's exact test.
* $p < .05$

Opt-in Web Panel A and Opt-in Web Panel B yield significantly higher estimates of agreement across all five EA questions than the ANES panel. The average difference is 14 points, with a high of almost 18 points for EA1 between Opt-in A and ANES.

Furthermore, additional analysis of the data by us reveals significant differences between the non-probability panels and the ANES panel across age and racial groups, gender, and education levels (see Table 2).

Table 2: Number of EA attributes with response distributions found to be statistically different from ANES Web Panel.

	Knowledge Panel	Opt-in Web Panel A	Opt-in Web Panel B
All Respondents	2	5	5
Age			
Under 35 yrs. old	0	5	5
35-55 yrs.	3	5	5
Over 55 yrs. old	4	3	4
Race/Ethnicity			
White	2	4	5
African American	0	5	5
Hispanic	1	4	5
Other	1	2	4
Female	2	5	5
Education			
High school diploma or less	1	5	5
More than high school diploma	2	5	5

Difference compared to ANES Web Panel uses Fisher's exact test.

The size and robustness of the differences found among the opt-in panels compared to the probability-based panels mean that these same EA questions can be used as calibration benchmarks for the majority of Internet survey studies that need to blend probability and non-probability samples. Adding these five EA questions to an online survey (administered as a single grid presentation) adds minimal cost or time to the survey. This saves the researcher from having to identify a different set of questions for each project requiring calibration—questions that are likely untested in their effectiveness to differentiate the two types of sample sources. Next, we demonstrate Knowledge Networks' calibration technique and how it makes use of these EA questions.

Calibration Approach

Step 1

There are three steps to the calibration method. The first step requires weighting only the probability portion of the sample. This is fundamentally a post-stratification raking procedure using a defined set of geographic and demographic variables[1]. Each panel member has an associated base weight that adjusts for selection probability and other sample design features

1 Age, gender, race, Hispanic ethnicity, language proficiency (among Hispanic respondents), Census region, metropolitan status, education, household income, homeownership, and Internet access.

corresponding to their respective recruitment cohort. That base weight (bw_i^{KP}) is the starting weight used in the post-stratification raking procedure. Thus, the adjustment factor (w_i^{KP}) with the base weight constitutes the post-stratification weight (W_i^{KP}). The sum of the post-stratification weights is represented as follows:

$$\sum_{i=1}^{n^{KP}} W_i^{KP} = \sum_{i=0}^{n^{KP}} (bw_i^{KP} \; x \; w_i^{KP})$$

where:
bw_i^{KP} = KnowledgePanel member base weight
w_i^{KP} = KnowledgePanel member post-stratification adjustment factor

To control for outlier weights, the distribution of W_i^{KP} is conservatively trimmed (Windsorized) at approximately the 1st and 99th percentile (using the most logical corresponding cut-off point displayed by the distribution). This weighted and trimmed probability sample now provides the benchmarks for the next step.

Step 2

The second step is to combine the weighted probability sample with the unweighted opt- in panel sample. These combined cases are then weighted overall to the probability sample's benchmarks from the previous step. Again, a post-stratification raking procedure is used. In this step, the starting weight for the probability-based cases is W_i^{KP}. However, because the opt-in cases have no known selection probability, we assign a value of 1.0 as their base weight (bw_i^{opt}) and use that as their starting weight in this step's post-stratification procedure. Using all the cases, the post-stratification adjustment factor (w_i^{All}) is a multiplier with each case's relevant starting weight to produce a final combined weight (W_i^{All}). The sum of the combined post-stratification weights for all cases is represented as follows:

$$\sum_{i=1}^{N^{All}} W_i^{All} = \sum_{i=1}^{n^{KP}} (W_i^{KP} \; x \; w_i^{All}) + \sum_{i=1}^{n^{opt}} (bw_i^{opt} \; x \; w_i^{All})$$

where:
w_i^{All} = All cases combined post-stratification adjustment factor
bw_i^{opt} = 1.0 as the opt-in panel base weight

Again, to control for outlier weights, the distribution of W_i^{All} is conservatively trimmed at approximately the 1st and 99th percentile (using the most logical corresponding cut-off point displayed by the distribution). This weighted and trimmed sample is now the "blended" sample to be evaluated in the next step.

Step 3

In the third step, we compare the answers from the five early adopter questions (EA1- EA5) between the probability sample from step 1 to the answers from the blended sample from step 2. The proportions used are the combined top two boxes of agree and strongly agree. Additionally, we identify at least three key questions as indicator study variables (SV1-SV3) to further assess bias introduced into the blended sample by the non- probability cases. The bias can be observed as differences in the point estimates. The assumption is that the point estimate of the KN Panel sample is the true reference. If the combined cases do not appreciably alter the point estimates of the study variables, we might possibly conclude that no calibration is necessary. However, if differences are observed then we proceed with a calibration step. Also, since we are only looking at a limited set of study variables yet we observe differences in the EA questions, it may be prudent to proceed with a calibration step in consideration for other possible (unobserved) differences.

Figure 1 is an example of a comparison of EA questions and three study variables where there are 105 probability panel cases (called "KN Panel" in Figure 1) and a total of 174 probability plus opt-in panel cases (called "Blended" in Figure 1). The point estimates are all moved higher for the five EA questions and also the three SV questions when the Blended data are compared to the KN Panel data. Because we observe these differences, we proceed to perform calibration in the next step.

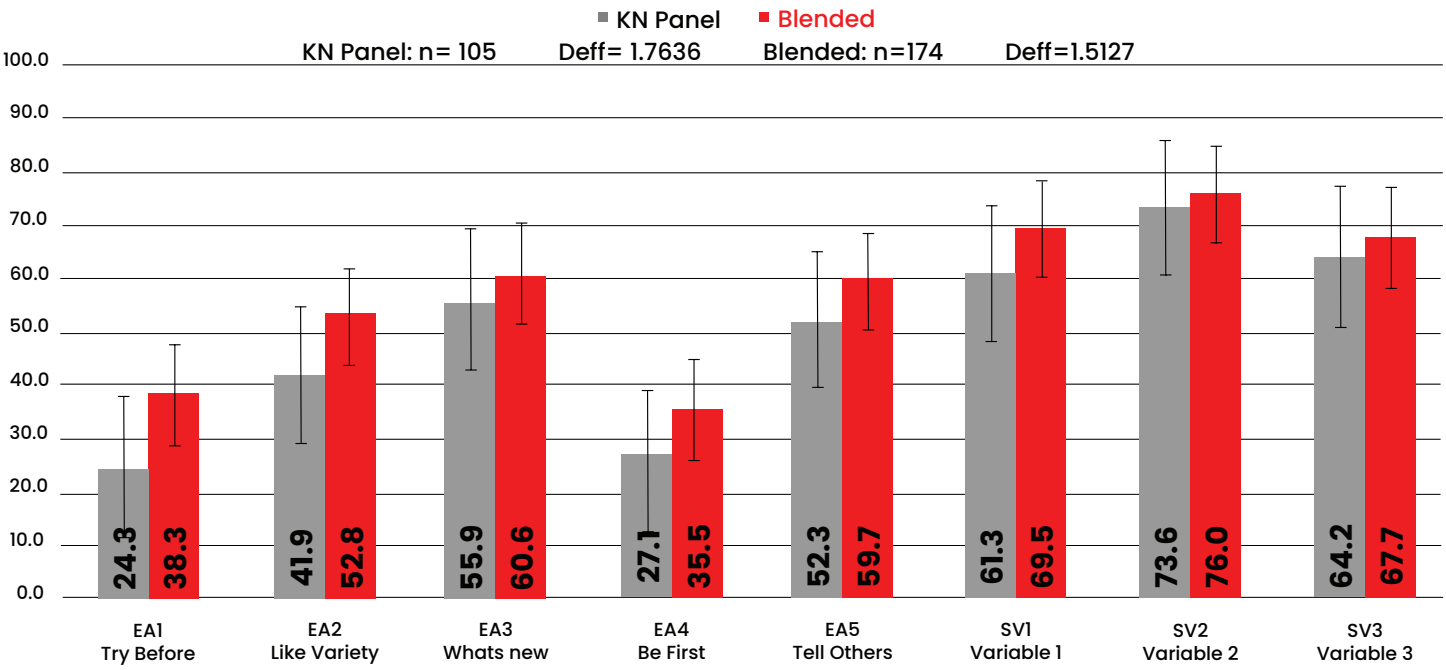

Figure 1. Example of observed differences among EA and SV questions between KN Panel and Blended samples with no calibration performed.

Step 4

In this step we select some minimum number of EA questions to include in the raking procedure carried out in Step 2. Generally, we try not to include more weighting variables than necessary to achieve a reduction in observed differences. We tend to select the EA questions that show the greatest differences. However, by choosing the fewest number, we sometimes need to repeat the Step 2 process, adding an additional EA question to further reduce differences. In the example shown in Figure 1, we selected EA1, EA2 and EA4 to be included in the Step 2 raking procedure. The results are shown in Figure 2 where the point estimates among the EA questions and among the three study variables are now more closely aligned. The goal is always to minimize bias introduced by the non- probability cases and more specifically minimize differences among the study variables.

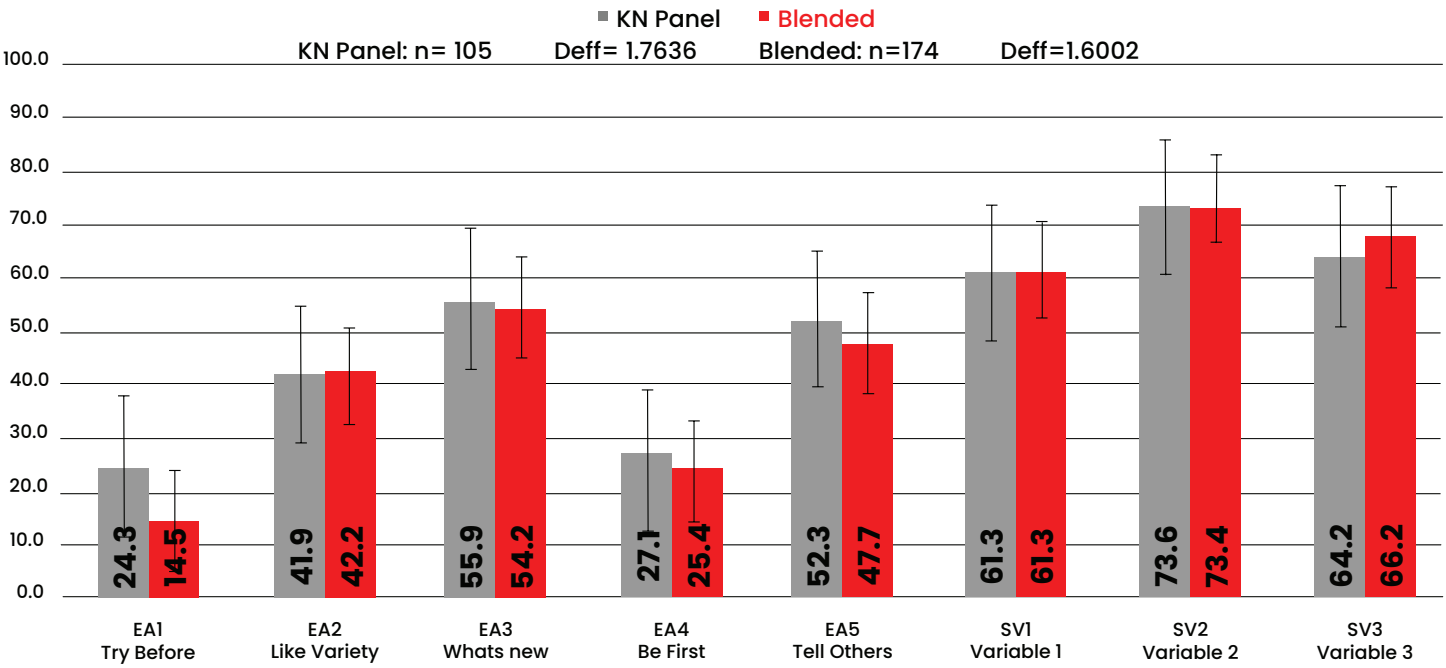

Figure 2. Example of reduced differences among EA and SV questions between KN Panel and Blended samples after calibration is performed using EA1, EA2 and EA4.

Evaluation of Calibration Approach

The calibration weighting approach presented above is easy and efficient for analysts to complete quickly for any project. We have used the five EA questions and three study questions to do a rapid visual assessment. However, the question that remains is whether this way of blending data from probability and non-probability samples yields estimates that are as good as or better than alternative approaches. And, for a more robust quantitative evaluation, we would want to examine a wider array of study questions. To do this, we compare estimates produced for 13 attitudinal variables using 611 cases from a probability-based sample and 750 cases from an opt-in sample that come from a study of attitudes toward smoking in a large mid-western state.

The probability-based sample and non-probability opt-in sample were examined in five different sets, three of which handled the blending process differently. Set 1 weights the 611 probability-based cases to Census population benchmarks (e.g., age, gender, education, etc.). The study questions from Set 1 will be used as the reference or "gold standard" against which to compare the other sets. Set 2 weights the 750 opt-in cases to the same Census population benchmarks, a likely approach for researchers who choose to use only opt-in samples. Set 3 blends the weighted probability-based cases with the *weighted* opt-in cases and then *re-weights all* to the same benchmarks in Set 1 using no calibration at all. Set 4 blends weighted probability-based cases with *weighted* opt-in cases then re-weights all to the Set 1 benchmarks this time using EA questions for calibration. Finally, Set 5, our recommended approach, blends the weighted probability- based cases with *unweighted* opt-in cases then re-weights all to the Set 1 benchmarks using EA questions for calibration. The 13 attitudinal variables used for comparison include 11 5-pt. agree/disagree Likert scales and 2 dichotomous agree/disagree questions. Calibration for Sets 3-5 was done using questions EA1, EA3, and EA5.

We evaluate the quality of estimates in five different ways. First, we treat estimates from the weighted probability-based sample only (Set 1) as unbiased point estimates and then compare them to estimates obtained from each of the four other sets of data. We report the absolute percent differences as the average absolute error across the 13 measures. We also report how many of the 13 attitudinal measures in each of the four other datasets differ from Set 1 by more than 2 percentage points. Next, as a quality metric, we report the design effect (Deff) of each dataset This is a measure of the variance in the weights. The greater a sample deviates from the benchmarks, more extreme weights are necessary to correct the distribution. This is measured as a larger Deff and consequently reduces the study's effective sample size thus lowering the value of the sample to make stable, generalizable estimates.

Two additional quality metrics calculate the bias and mean squared error (MSE) of estimates produced by each set. Bias and MSE quantify the differences between values implied by an estimator and the true values of the quantity being estimated. True bias is not known, but we can estimate it using the following equation from Ghosh-Dastidar et al. (2009):

$$\hat{\varepsilon}^2 = \max\left(0, (\bar{x}_{Set\ 1} - \bar{x}_{Set\ x})^2 - \frac{s^2_{Set\ 1}}{n_{Set\ 1} - 1} - \frac{s^2_{Set\ x}}{n_{Set\ x} - 1}\right)$$

The above equation, using Set 1 as the reference estimate, subtracts from the observed square difference the quantity expected from sampling variance alone, leaving an estimate of squared bias. It further ensures a minimum estimate of zero. MSE incorporates both the variance of the estimate (a measure of precision) and its estimated bias (Ghosh-Dastidar et al. 2009). It can be calculated as follows:

$$MSE_{Set\ x} = \hat{\varepsilon}^2 + \frac{s^2_{Set\ x}}{n_{Set\ x} - 1}$$

Results

The average absolute error when compared to Set 1 ranged from 5.3% for Set 2 to 1.3% for Set 5 (see Table 3). Furthermore, the number of estimated items with an absolute error of 2 or more percentage points was 12 out of 13 for Set 2 and only 3 out of 13 for Set 5. Calibration without pre-weighting the opt-in cases first (Set 5) performed slightly better than calibration with pre-weighting the opt-in cases first (Set 4), with a slightly lower average absolute error (1.3% vs. 1.7%) and less than half the estimated items with an absolute error of 2 or more percentage points (3 items vs. 7 items). The design effect is also lowest for Set 5 compared to all other estimates except Set 1, the probability-based sample only.

Table 3: Evaluation of calibration technique on 13 selected attitude items.

	Set 1*	Set 2	Set 3	Set 4	Set 5
Item 1	51.0%	60.1%	54.8%	53.6%	53.6%
Item 2	69.5	74.5	71.3	70.8	70.7
Item 3	49.6	43.9	46.8	47.1	47.9
Item 4	54.3	48.0	51.5	52.0	53.0
Item 5	48.1	44.7	46.4	47.0	47.7
Item 6	46.4	42.1	43.8	44.0	44.6
Item 7	41.4	48.6	44.5	43.7	42.9
Item 8	44.4	45.4	44.9	44.1	44.2
Item 9	42.5	46.6	43.8	43.4	43.7
Item 10	63.6	71.2	67.3	66.2	65.7
Item 11	53.1	59.3	55.8	55.2	55.2
Item 12	31.8	34.3	33.1	32.5	32.6
Item 13	36.6	42.6	38.3	37.0	37.1
Number of cases in sample	611	750	1,361	1,361	1,361
Average Absolute Error	--	5.3%	2.3%	1.7%	1.3%
No. of items with error of 2 or more percentage points	--	12	7	7	3
Deff	1.872	3.480	2.155	2.240	2.095
Average Estimated Bias	--	25.579	2.056	0.190	0.064
Average Estimated MSE	3.937	28.741	3.816	1.950	1.826

***Notes:** **Set 1** is weighted probability-based sample only (reference data).

Set 2 is weighted non-probability opt-in sample only.

Set 3 blends weighted probability-based samples with *weighted* non-probability opt-in sample and then *re-weighted* using no calibration.

Set 4 blends weighted probability-based sample with *weighted* non-probability opt-in sample then *re-weight* to benchmarks using calibration.

Set 5 blends weighted probability-based sample with *unweighted* non-probability opt-in sample then *re-weight* to benchmarks using calibration (recommended approach).

Set 2, the opt-in sample only, had the largest MSE (28.741), with much of its MSE composed of bias (25.579; see Table 3). The next largest MSE was for Set 1 at only 3.937, approximately one-seventh of the MSE of Set 2. Calibration without first pre- weighting the opt-in cases (Set 5) yielded the lowest average MSE (1.826), which is less than half that of Set 1 (probability-based sample only). The average estimated bias of Set 5, the lowest of all the other sets, is about three times less than the next lowest set, Set 4.

Conclusion

National probability-based Internet panels have been limited by their sample size from being useful in studies of small geographic areas or rare incidence phenomena; opt-in Internet panels, on the other hand, have a reputation for yielding low quality and biased estimates. This paper demonstrates a calibration technique to overcome the limitations of both types of panel data by combining the samples in a way that is relatively easy and successfully minimizes bias in the resulting larger combined sample. We demonstrate through a quantitative evaluation that the estimates obtained from the calibration approach laid out in this paper result in the smallest average absolute error, lowest estimated bias, and smallest average mean squared error than other data combination techniques. It is unnecessary to compute post-stratification weights for the opt-in sample prior to calibration.

The results of the evaluation done in this paper are also consistent with earlier research which states that estimates from non-probability samples are often substantially biased, even after quota sampling and post-stratification weighting. In fact, the opt-in sample only (Set 2) yielded estimates with the highest average squared error, highest number of items with a difference of 2 percentage points or more, largest average bias, and highest average mean squared error. Opt-in samples alone are not a viable data collection solution at this time for studies that require accuracy and generalizable results.

Calibration requires researchers to have measures at hand that can differentiate probability-based samples from opt-in samples. This paper demonstrates a series of five questions related to the early adoption of new products and technology that appear to reliably distinguish between the two types of sample respondents both as a whole and within many specific demographic categories. Knowing the reliability of these questions *a priori* means that they can be added to any questionnaire in which calibration may be necessary, with little risk of failure to distinguish the panels.

This calibration technique that uses early adopter measures serves the rapid data turnaround required of many research projects with only a nominal increase in effort. Moreover, the calibrated combination of data does not appear to add significant bias or variance to the estimates it yields. However, continued research is necessary to better understand the underlying statistical implications of using a calibrated dataset for reliable generalizable estimates. Also, more work on early adopter measures as to the limits of their applicability is encouraged. We believe that the calibration approach we describe is a viable methodology for combining probability and non-probability samples derived from Internet panels.

References

AAPOR Opt-in Online Panel Task Force (2010). *AAPOR Report on Online Panels.*

American Association of Public Opinion Researchers.

American Association of Public Opinion Research (2011). *Opt-in Surveys and Margin of Error*. At: http://www.aapor.org/Content/aapor/Resources/PollampSurveyFAQ1/OptInSurveysandMarginofError/default.htm

Bethlehem, J., & Stoop, I. (2007). Online Panels--A Paradigm Theft? In T. Trotman, T. Burrell, L. Gerrard, K. Anderton, G. Basi, M. Couper, et al. (eds.), *The Challenges of a Challenging World: Developments in the Survey Process* (pp. 113-131). Berkeley, UK: Association for Survey Computing.

Bourne, F. S. (1959(2001)). The Adoption Process. In M. J. Baker (ed.), *Marketing: Critical Perspectives on Business and Management.* New York: Routledge.

Chang, L., & Krosnick, J. A. (2009). National Surveys via RDD Telephone Interviewing vs. the Internet: Comapring Sample Representativeness and Response Quality. *Public Opinion Quarterly* , 73 (4), 641-678.

Couper, M. P. (2000). Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly* , 64, 464-494.

de Leeuw, E., & Van der Zouwen, J. (1988). Data Quality in Telephone and Face-to-Face Surveys: A Comparative Meta-Analysis. In R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholss, & J. Waksberg (eds.), *Telephone Survey Methodology*. New York: Wiley.

Dennis, J. M., Osborn, L., & Semans, K. (2009). *Comparison Study: Early Adopter Attitudes and Online Behavior in Probability and Non-Probability Web Panels.* Available at http://www.knowledgenetworks.com/accuracy/spring2009/pdf/Dennis- Osborn-Semans-spring09.pdf

Dillman, D. A. (2000). *Mail and Internet Surveys: The Tailored Design Method.* New York: Wiley.

Fricker, R. D., & Schonlau, M. (2002). Advantages and Disadvantages of Internet Research Surveys: Evidence from the Literature. *Field Methods* , 14 (4), 347-367.

Ghosh-Dastidar, B., Elliott, M. N., Haviland, A. M., & Karoly, L. A. (2009). Composite Estimates from Incomplete and Complete Frames for Minimum-MSE Estimation in a Rare Population: An Application to Families with Young Children. *Public Opinion Quarterly* , 73 (4), 761-784.

Goyder, J. (1985). Face-to-Face Interviews and Mailed Questionnaires: The Net Difference in Response Rate. *Public Opinion Quarterly* , 49 (2), 234-252.

Guo, S., & Fraser, M. W. (2010). *Propensity Score Analysis: Statistical Methods and Applications.* Thousand Oaks, CA: Sage Publications.

Kott, P. S. (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors. *Survey Methodology* , 32 (2), 133-142.

Kypri, K., Stephenson, S., & Langley, J. (2004). Assessment of Nonresponse Bias in an Internet Survey of Alcohol Use. *Alcoholism: Clinical and Experimental Research* , 28 (4), 630-634.

Pasek, J., & Krosnick, J. A. (2010). *Measuring Intent to Participate and Participation in the 2010 Census and Their Correlates and Trends: Comparisons of RDD Telephone and Non-Probability Sample Internet Survey Data.* Washington, DC: U.S. Census Bureau.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. J*ournal of the American Statistical Association* , 102, 75-83.

Rueda, M., Martinez, S., Martinez, H., & Arcos, A. (2007). Estimation of the Distribution Function with Calibration Methods. *Journal of Statistical Planning and Inference* , 137 (435-448).

Sarndal, C.-E. (2007). The Calibration Approach in Survey Theory and Practice. *Survey Methodology* , 33 (2), 99-119.

Schonlau, M., Van Soest, A., & Kapteyn, A. (2007). *Are 'Webographic' or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring*. RAND Corporation. RAND Corporation.

Skinner, C. (1999). Calibration Weighting and Non-Sampling Errors. *Research in Official Statistics* , 2, 33-43.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., et al. (in press). Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples. *Public Opinion Quarterly* .