LUTH
research

Intelligence from knowledge.™

An Empirical Approach to Correct Self-Selection Bias of Online Panel Research

Presenter:
Lisa Luth, Chief Knowledge Officer
Luth Research

2008 CASRO Panel Conference

Non-probability sampling has long been considered a critical risk to the validity of research results, especially when the goal is to generate an accurate projection or generalization for the entire population. The fundamental difference between non-probability sampling and probability sampling is the former lacks random selection – the members of the target population are not being given an equal opportunity to be selected. Thus our sample cannot provide researchers the opportunity to calculate the probability of the occurrences of the findings within a given population. Non-probability sampling takes many forms, but is primarily considered either accidental or purposive (http://www.socialresearchmethods.net/kb/sampnon.php). Accidental non-probability sampling refers to the convenience sampling employed by media outlets for conducting polls for news or entertainment purposes, or by academic research that typically use college students. Purposive non-probability sampling is often seen in the form of mall intercept sampling, snowball sampling, expert sampling, and quota sampling.

Compared to the traditional random-digit-dialing telephone interviewing method, online panel research has been generally criticized as relying on self-selection recruitment, and thus is based on non-probability sampling. The risk associated with non-probability sampling is not unique to online panel research, but has been amplified by the hard-to-control online panel recruiting process. Even though online panel has become a more widely accepted research methodology, new and additional challenges emerge in ensuring data quality and sound sampling practices. In addition to the leading initiatives by ESOMAR, CASRO, and ISO in establishing standards and guidelines to promote online research quality, the Advertising Research Foundation has also recently formed its new Online Research Quality Council to specifically further debate and development in formulating relevant metrics to measure quality. Out of these efforts, a few particular concerns for current online research quality relate to the existing sampling practices employed by many online panel and online research companies (Tomei, 2007; comScore report at CASRO cited by Sparrow, 2007):

1. To capitalize on the speed and cost benefits of the online methodology, clients and suppliers may compromise the checks and balances implemented in traditional methodologies in ensuring a representative sample.
2. Online panel recruiting is based on people's willingness to participate in research activities; financial incentives are in many cases a significant influence on people's decision in becoming a panel member or participating in online research.
3. With the online panel approach, a relatively select group of active respondents may account for the majority of the survey participation

These concerns again point to the self-selection nature of most online panels. While there are extensive discussions and acknowledgements about this self-selection limitation, insufficient efforts have been made in the industry to develop and debate logical remedies to this leading methodological issue for online research.

Researchers have very few systematically developed, empirically tested corrective mechanisms to address the self-selection bias embedded in online panel samples. Several more commonly known solutions include establishing quotas to control for primarily demographic distributions of online sample, weighting the ending online sample to reflect the target population distributions. A few panel companies have also made efforts to balance their panel populations to better represent the overall population. For example, Knowledge Networks challenges the fundamental problem with general online samples, and has developed a system to include non Internet population segments by providing them with Internet access in an effort to correct this skew towards Internet usage (Pineau & Slotwiner, 2003). Polimetrix provides another method by conducting sampling by location. In other words, they will include their online panel members into an online sample if the panel members exactly match or are located near the members of a sample drawn in the traditional random way.

The developments beyond simple quota setting discussed in the above are applaudable. However, these examples tend to be difficult for replication and at times can induce tremendous investments from the undertaking company. As we seek to advance our understanding and practices of online panel sampling as a whole, one cannot help but ask if there is a methodology that individual research panel companies can all utilize, cross-review, and improve on as we continue to use online panel as an integral part of the available research tools. In our perspective, propensity scoring can be used as such a method.

Luth Research revisited the option of developing a propensity score system as a way to normalize the self-selection bias in panel sampling. The propensity score approach was first used by Rosenbaum & Rubin (1983) in observational research to balance treatment and control subjects. It is defined as a "conditional probability of exposure to a treatment given observed covariates." In the case of online research panel, the propensity score developed in this paper represents the conditional probability of being a panel member given specific observed covariates. Using propensity score as a balancing approach is not a brand new concept to the market research industry, especially in survey research. It has been used by some research companies as well as academic researchers to correct the discrepancies between results from a Web-based survey (typically relying on active Web users) and those from the expected target population (e.g., Harris Interactive; Danielsson, 2000; Schonlau, et al., 2003, 2007).

Methodology

As the first step, we conducted a telephone study using random digit dialing and an online study using online sample randomly drawn from Luth Research's SurveySavvy™ panel. The RDD study included 800 completed interviews and the online study included 144 completed interviews. Both studies used the same questionnaire, which included a set of demographic variables and a set of attitudinal questions. The demographic variables were gender, age, ethnicity, income, education, marital status. The attitudinal questions measured to what degree each of the following statements describes the respondent:

1. I am one of the first people to use or buy a new product
2. I tend to keep opinions to myself
3. I consider my time very valuable
4. My schedule is always busy
5. Like to make my voice heard
6. Financial rewards are usually important to me
7. My friends would say I am technology savvy

The RDD telephone survey is the proxy to conducting a true representative survey among the overall population. The general argument is that telephone penetration is still significantly higher than Internet penetration; therefore it is reasonable to expect a RDD telephone survey to have a better chance to represent the population (Schonlau, et al., 2004). In addition, thanks to the RDD approach, a telephone survey provides the inferential probability that researchers need to make extrapolation to the population.

The question regarding which attitudinal questions to include for effective comparison has been left open so far in related literature. The type of attitudinal questions and the number of attitudinal questions available for inclusion are of no limitation. While there is no consensus on the criteria to develop these attitudinal questions, the decision typically hinges on what makes most sense in establishing the potential differences between the sample and the target audience involved in the survey or the survey subject (e.g., Sparrow, 2007). An important purpose of existing and recent research efforts delving into propensity score development is actually to provide empirical investigation on which attitudinal questions may present more fruitful and meaningful explanation for the differences between the sample and the target population (Lee, 2004, 2006; Schonlau, 2007). In this paper, we looked at the general practices of online panel companies as well as the common respondent participation issues before we selected the aforementioned attitudinal questions. First, broadly speaking, we found that most panel companies use at least one form of financial incentive to motivate people to join and participate in an online panel. Secondly, as online panel is still a relatively new idea and requires some level of technology savvy, we expected people joining a panel to be more open to new ideas and possess a higher level of technology proficiency. Another factor generally understood as a benefit of online panels is the convenience of participating in research, which may be more appreciated by people who value their time. Lastly, there is an expected difference between panel members and the general population in the extent to which they are willing to voice or withhold their opinions. The above four general areas have found support for potential concerns in relation to the selection critique that is often raised for sampling issues in research literature (Arzheimer & Klein, 1999; Schillewaert & Meulemeester, 2005).

In the next step, the data from both surveys were used to form a single data set to allow for a logistical regression analysis, which is commonly employed to develop the propensity scores. The logit model is:

$$\Pr(T_i = 1 | X_i) = \frac{e^{\lambda h(x_i)}}{1 + e^{\lambda h(x_i)}}$$

Where T*i* is the SurveySavvy panel membership status and h(x*i*) is made up of linear and/or higher order terms of the covariates. We define the propensity score as the conditional probability of SurveySavvy panel membership given a vector of measured covariates – the seven attitudinal questions.

The propensity scores were then calculated using the following:

$$\ln\left[\frac{\Pr(T_i = 1 | X_i)}{1 - \Pr(T_i = 1 | X_i)}\right] = \lambda h(x_i)$$

The results of the logistical regression indicated the following variables made a significant contribution to predicting the likelihood for a respondent to be in the SurveySavvy panel:

Attitudinal variables:

I am one of the first people to use or buy a new product

I tend to keep opinions to myself

I consider my time very valuable

My schedule is always busy

Financial rewards are usually important to me

My friends would say I am technology savvy

Table 1.0 Coefficients of attitudinal variables in logistic regression.

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
I am usually one of the first people among my family and friends to use or buy a new product that is coming on the market.	-0.239	0.076	9.862	1	0.002	0.787
I tend to keep my opinions to myself.	0.250	0.070	12.812	1	0.000	1.284
I consider my time very valuable.	0.262	0.106	6.164	1	0.013	1.300
My schedule is always busy.	0.183	0.080	5.287	1	0.021	1.201
I like to make my voice heard.	-0.046	0.093	0.241	1	0.624	0.955
Financial rewards are usually important to me.	-0.262	0.085	9.466	1	0.002	0.770
My friends would say I am technology savvy.	-0.257	0.078	10.863	1	0.001	0.773
Constant	-1.345	0.373	13.020	1	0.000	0.260

Propensity scores were generated from the logistical regression model. The next step is to calculate a proper weight scheme based on the propensity scores to remove or minimize the differences between the online SurveySavvy panel members and the RDD telephone respondents. The existing research literature points to three general ways of developing this weight scheme: inversed propensity scores, propensity score stratification, and conventional multivariate modeling (Glynn, Schneeweiss & Sturmer, 2006; Schonlau et al, 2007). However, there is no single most superior method among these options and in fact, Glynn et al (2006) concluded “overall, the alternative implementation of propensity score methods gave estimates similar to each other and to conventional multivariate models”. For this paper, we used the propensity score stratification method to develop the weight scheme. The stratification method has been found to have a better ability to preserve the sample count or the analysis cell count (Hazelwood et al, 2007).

Using the stratification method, we divided the total number of estimated propensity scores in the combined data set including both the RDD and SurveySavvy respondents into five bins with each containing 20% of the total respondents. The number of bins derived is dependent on how well the bin stratification reduces the differences between the two samples. T-tests were performed to test the distributions of the stratified bins until a satisfactory level of balance was achieved. Five bins were determined as the optimal level of stratification for this research effort. The distribution of the RDD phone respondents across the five bins was used as the target distribution and the distribution for SurveySavvy panel respondents was weighted to match the target RDD distribution across the five bins. An overall weight was then derived for each respondent in the SurveySavvy data set for correcting the panel selection bias.

The weight scheme resulted from the above analysis became the corrective mechanism to balance the differences between the SurveySavvy panel and the general population. We adopted Schonlau et al's approach in examining whether the weight scheme was able to correct the panel bias as expected. This approach looked at two measurements: 1) how closely the weighted means of the attitudinal variables match against the results of the same variables for the target population; 2) the changes in effect size before and after applying the weights.

As indicated in Table 2.0 below, there is no significant difference between the adjusted means of the attitudinal variables for the SurveySavvy sample and those for the RDD sample. The effect sizes for the attitudinal variables after the weighting adjustment were significantly reduced.

Table 2.0. Comparison of unadjusted and adjusted attitudinal measurements.

Variables	**Values**	**Phone**	**SurveySavvy Panel**		**Effect Size (Odds Ratio/Cohen's D)**		**Effect Size Difference**
			Unadjusted	Adjusted (Webographic)	Phone vs. Unadjusted SurveySavvy Panel	Phone vs. Adjusted SurveySavvy Panel (Webographic)	|Adjusted|-|Unadjusted|
One of the first people to use or buy a new product	Mean	3.29	2.65	3.47	0.390	-0.108	-0.282
Tend to keep opinions to myself	Mean	3.21	3.68	3.24	-0.272	-0.016	-0.257
Consider my time very valuable	Mean	1.57	1.75	1.62	-0.129	-0.035	-0.094
My schedule is always busy.	Mean	2.08	2.40	2.06	-0.198	0.011	-0.187
Like to make my voice heard	Mean	2.25	2.04	2.30	0.140	-0.029	-0.111
Financial rewards are usually important to me.	Mean	2.30	1.97	2.42	0.201	-0.070	-0.131
My friends would say I am technology savvy.	Mean	3.11	2.47	3.16	0.386	-0.025	-0.361

Note: The phone sample and the unadjusted SurveySavvy Web sample were weighted to reflect their respective target population distributions.

Validation

After we verified the propensity score approach that we applied achieved the balancing effect that we had desired in the test data, we proceeded to validate this approach by conducting an additional online study using SurveySavvy panel sample.

With a total of 802 completed interviews, the second online study included all the attitudinal questions used in the initial study as well as several factual questions for which overall population data are available for benchmark comparison. The additional factual questions were whether or not the respondent smokes and the political party affiliation the respondent is associated with. The smoking question was chosen primarily for the purpose of having a population behavioral parameter that is readily available for comparison. Because of the lack of general national attitudinal benchmarks, the political party affiliation question was used as a shortcut to gauge the current attitudinal inclinations of the population.

Following the same approach described in the above, both the online study attitudinal data and the telephone RDD attitudinal data were entered into a logistic regression from which the propensity scores were derived. The same stratification method was applied and five bins were determined to best normalize the differences between the online data and the telephone RDD data. An overall weight was calculated and used to adjust for differences between the panel data and the RDD data.

The results showed that the mean scores for all attitudinal questions were closely matched between the adjusted panel sample and the RDD sample. In addition, the party affiliation distribution for the adjusted panel sample was much more aligned with that of the general population (reported by the American National Election Studies Partisanship and Evaluation of the Political Parties 2004). However, it is interesting to note that the result for the smoking question did not improve after the adjustment. The proportion of smokers remains significantly higher in the panel sample than in the general population.

Table 3.0 Comparison of unadjusted and adjusted measurements in validation study

Variables	**Values**	**Population /Phone**	**SurveySavvy Panel**		**Effect Size (Odds Ratio/Cohen's D)**		**Effect Size Difference**
			Unadjusted	Adjusted (Webographic)	Phone vs. Unadjusted SurveySavvy Panel	Phone vs. Adjusted SurveySavvy Panel (Webographic)	|Adjusted-1| - |Unadjusted-1| or |Adjusted|- |Unadjusted|
Smoker	Yes	20.8%	38.2%	37.1%	0.425	0.445	-0.020
Party affiliation	Republican	41.0%	29.5%	40.7%	1.661	1.012	-0.648
	Democrat	49.0%	36.9%	41.5%	1.643	1.354	-0.289
	Independent	10.0%	24.6%	12.3%	0.341	0.792	-0.452
	Other	0.0%	9.1%	5.5%	0.000	0.000	0.000
One of the first people to use or buy a new product	Mean	3.29	3.06	2.94	0.142	0.217	0.075
Tend to keep opinions to myself	Mean	3.21	2.53	3.15	0.422	0.040	-0.381
Consider my time very valuable	Mean	1.57	4.08	2.18	-1.828	-0.413	-1.414
My schedule is always busy.	Mean	2.08	3.65	2.88	-1.022	-0.511	-0.511
Like to make my voice heard	Mean	2.25	3.61	2.62	-0.902	-0.232	-0.670
Financial rewards are usually important to me.	Mean	2.30	3.85	2.65	-0.997	-0.203	-0.794
My friends would say I am technology savvy.	Mean	3.11	3.45	3.39	-0.208	-0.165	-0.042

Note: The phone sample and the unadjusted SurveySavvy Web sample were weighted to reflect their respective target population distributions.

Discussion

Online panel recruiting practices will not change overnight and there will always be influential elements that contribute to the relatively limited availability of sample for market research. Our investigation in using propensity scores developed based on attitudinal questions as a corrective mechanism for the online panel selection bias is intended to promote further discussion about establishing effective post hoc procedures to ensure panel data quality. This paper has brought attention to three areas when developing propensity scores for the purpose of correcting online panel sampling error: 1) Using attitudinal questions, 2) treating each panel individually, and 3) developing an open rather proprietary mechanism.

Using attitudinal questions

Propensity score development and typical weighting procedure have often used demographic variables and factual questions such as ownership or usage of certain services or products in comparing the treatment group and the control group. While these measurements are usually effective in identifying the differences between the two samples, we are convinced that it is necessary to uncover relevant attitudinal questions that will help understand the contrasts of the two groups because people manifesting similar behaviors may be driven by varying psychological traits and values. When we start to integrate these relevant attitudinal questions into the weighting procedure, we expect to see a higher level of precision in making projections to the overall population.

In our perspective, determining the relevant attitudinal questions is an on-going effort. Through this research effort, we have identified that perceived value of time, motivation by financial rewards, inclination to keep opinions to oneself, likelihood to adopt new technology are four areas where SurveySavvy panel members are different from the overall population, and should receive correction to normalize the differences. However, it is reasonable to speculate that there are many more aspects in which the panel population distinctively differs from the general population. To best capture, monitor and account for these variances, we recommend a set of 10 to 20 questions be developed and reassessed on their relevance annually based on the changes in the composition of the online panel, the online population and the overall population. These questions can include economic, political and lifestyle attitudes.

Furthermore, our research discovered the attitudinal questions related to the four areas (perceived value of time, motivation by financial rewards, inclination to keep opinions to oneself, likelihood to adopt new technology) were not good "balancing" factors in projecting smoking behaviors, and thus compelled us to think if behavioral questions (e.g., smoking) need to be built into the repository of measurements beyond demographic and attitudinal questions to best understand and account for the differences between an online panel and the overall population. Future studies will need to further examine the benefit of the integration of demographic, attitudinal and behavioral measurements in developing effective propensity scores for online panel members.

Treating each panel individually

While our research is focused on Luth Research's SurveySavvy panel, we expect that existing online research panels vary in the extent to which each differs from the overall population due to their different recruiting methods. Therefore, some attitudinal questions may be more effective in accounting for a specific panel's differences from the population than others.

Similar to panel demographic distributions, attitudinal inclinations are also unique characteristics of a panel. As the market research industry is promoting transparency in understanding each panel's demographic dispositions, weakness and strengths in targeting specific demographic groups, it is important to also recognize the impact of the underlying psychological composition of the panel on survey sampling. This recognition calls for not only publishing and revealing significant attitudinal inclinations represented by the various panels, but also developing customized approaches to mitigate these "biases" for specific panels.

Developing an open method rather than a proprietary mechanism

To date, most corrective approaches for sampling error or panel quality issues are proprietary. While a proprietary mechanism is best in protecting business know-how and ensuring certain business interests, we argue that developing propensity scores for the purpose of correcting selection bias in panels will be better served by having an open method, shared development efforts and industry-wide participation. Two main factors contribute to this reasoning. First, panel quality issues, especially the selection bias, have been elevated to be a top concern for using online research in general. This is not pertinent to only one specific panel. Rather, it is a concern for all panels. Secondly, the number of attitudinal dimensions where the panels may differ from the population and need to be accounted for is expected to be large. Currently, there is a lack of discussions and validations on which attitudinal questions are most effective in normalizing the differences between a panel and the overall population. A pragmatic solution to finding the most relevant questions to account for the selection bias is to engage most if not all panels in voluntarily identifying the questions that are most effective to them, some of which could be significant to all panels. A joint effort by the panel providers will help advance researchers' trust in online panel research as a legitimate research methodology.

References

Arzheimer, K., & Klein, M. (1999). The effect of material incentives on return rate, panel attrition and sample composition of a mail panel survey. *International Journal of Public Opinion Research, 11*, 368-377.

Danielsson, S. (2000). The propensity score and estimation in nonrandom surveys – an overview. University of Linköping. Available online: http://www.statistics.su.se/modernsurveys/publ/11.pdf

Glynn, R., Schneeweiss, S, Stürmer, T. (2006). Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic and Clinical Pharmacology and Toxicology,* 98, 253-259.

Hazelwood, L., Mach, T., & Wolken, J. (2007). Alternative methods of unit nonresponse weighting adjustments: an application from the 2003 survey of small business finances. *Finance and Economics Discussion Series.*

Lee, S. (2004). Statistical estimation methods in volunteer panel Web surveys. Unpublished Doctoral Dissertation. University of Maryland, Joint Program in Survey Methodology.

Lee S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel Web surveys. *Journal of Official Statistics,* 22(2), 329-349.

Pineau, V & D. Slotwiner (2003). Probability Samples vs. Volunteer Respondents in Internet Research: Defining Potential Effects on Data and Decision-Making in Marketing Applications, Technical Paper, Knowledge Networks, California, USA.

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika,* 70(1), 41-55.

Schillewaert, N., & Meulemeester, P. (2005). Comparing response distributions of offline and online data collection methods. *International Journal of Market Research, 47(2),* 163-178.

Schonlau, M. et al., (2004). A comparison between responses from a propensity-weighted Web survey and an identical RDD survey. *Social Science Computer Review, 22(1),* 128-138.

Schonlau, M., Van Soest, A., & Kapteyn, A. (2007). Are Webographic or attitudinal questions useful for adjusting estimates from Web surveys using propensity scoring? RAND Population Research Center.

Sparrow, N. (2007). http://www.icmresearch.co.uk/white-papers/quality-issues-in-online-research.pdf

Tomei, R. (2007), Build a consensus, focus on quality. *Quirks,* November, 2007, 52-56.

Presenter:

Ms. Lisa Luth, Chief Knowledge Officer, Luth Research

With more than 12 years of experience in both qualitative and quantitative research, Lisa Luth oversees strategic design and execution of research projects as well as panel-based product development at Luth Research. Her expertise includes consulting on methodology selection, sampling plan development and analytical tool usage. In addition, Lisa is a highly respected, top-level moderator who has conducted hundreds of focus groups across the country with a wide variety of consumers and professionals.

Lisa previously held a position as Director of Research where she was responsible for managing the entire scope of market research functions for a chain of over 300 restaurants across the western and southern United States. Lisa also served as a Senior Manager of Product Research in PepsiCo's Taco Bell Division.

Lisa holds a Masters of Business Administration from Vanderbilt University.

About Luth Research

Since 1977, we have been offering our clients around the globe the kind of forward-thinking intelligence that moves their businesses ahead. From our innovative online panel, SurveySavvy™, to our creative combination of research methods by an experienced team, our clients count on us for excellence in market research.

For more information on Luth Research, visit our Web site at www.luthresearch.com or call a Luth Research sales professional now at 1.800.465.5884.