An Empirical Method to Establish Usability of Nonprobability Surveys for Inference

Robert D. Tortora[1], Ronaldo Iachan[1], Eric Miller[1]
[1]ICF International, 530 Gaither Rd, Rockville, MD 20850

Abstract

The most difficult issue to address under the "Fit for Purpose" (Baker, et al., 2013) approach to non-probability surveys is making point estimates that are statistically valid, that is, can be used for statistical inference. This paper describes an empirical methodology that moves survey research beyond the current state of affairs of comparing non-probability survey (NPS) estimates from a panel to probability survey (PS) estimates, the gold standard, to deciding when these NPS estimates can be used for statistical inference. The methodology calls for the development of an *a priori* set of decision rules which can be used to decide on the validity of the NPS. Three levels of use can be evaluated: 1) using the NPS for overall population estimates, 2) using the NPS estimates for subpopulation estimates and 3) using the NPS data for multivariate analysis. For each of these uses the NPS estimates are compared to the gold standard PS, say comparing confidence intervals and odds ratios from both surveys. The decision rule includes the *a priori* level of risk that can be tolerated for the NPS data to be used. The methodology goes one step further and illustrates an *a priori* decision rule where the need to make comparisons to a gold standard can be dropped and the NPS can stand alone and be used for inference on later occasions. Finally we illustrate our proposed empirical method by comparing data from a non-probability quota sample for the Los Angeles area with a RDD probability health survey of the same area.

Key Words: Non-probability Survey, Estimation, Inference, *a Priori* Decision Rule

1. Background

To date, most of the research surrounding the use of nonprobability survey (NPS) data has revolved around comparisons of estimates, see for example, Baker, Zahs, and Papa (2004), Berrens, Bohara, Jenkins-Smith, Silva, and Weimer (2002), Chang and Krosnick (2009), Couper (2007), Duffy, Smith, Terhanian, and Bremer (2005) and Iachan, Boyle, Harding (2016) with a probability survey (PS); see Lensvelt-Mulders, Lugtig and Hubregtse (2009); Bethlehem and Biffignandi (2010) for different evaluations of bias in a NPS; or Loosveldt and Sonck.(2008), Elliott (2009), Lee (2006), Lee and Valiant (2009), Valliant and Dever (2011) for methods to adjust a NPS to make it comparable to a PS. Perhaps not surprisingly there is a rather rich discussion of using quota samples for inference and also calculating sampling error for these surveys. Stephan and McCarty (1958) suggest using replicate estimates to calculate variance, Sudman (1965) suggests regular variance estimation works for multi-stage designs where the first stage is based on a probability design and later stages based on quotas. Smith (1983) examines conditions for ignoring non-random selection mechanisms and pays particular attention to poststratification and to quota sampling.

The most difficult part of the "fit for purpose" approach to using NPS data is when and how it can be used for statistical inference. While inference based on NPS has been widely accepted in the market research community the case has not been sufficiently made for acceptance in the social research community and the official statistics community. To broaden the acceptance of NPS in these latter two communities, this paper describes an empirically based methodology that moves beyond comparisons, bias evaluation and adjustment that evaluates the use NPS data for valid statistical inference.

For purposes of illustration, suppose that valid statistics are needed for an urban area over time and that there is insufficient budget to conduct a PS on the same topic on each occasion. Using an NPS' meets budget constraints and it so happens that there is sufficient budget to conduct one PS or that data from a contemporaneous PS is available. The PS can stand as the gold standard. The key questions then become 1) how can the NPS be judged as acceptable for inference when compared to the PS and 2) under what conditions can the NPS be used on later occasions for inference without having a PS for comparison? The approach is based on several key conditions: C1) the NPS is designed as a quota sample from a panel[1] where the demographic distribution of the panel for the target population is known; C2) the organization can outline beforehand what the NPS will be used for; C3) the organization can define *a priori* decision rule(s), inspired by the ASPIRE survey evaluation process (Bergdahl et al.,2014) that, if satisfied, will indicate that the NPS data is equivalent to the PS and is therefore acceptable to be used as described in C2 and C4) if on the first occasion the NPS is deemed acceptable, then if the change in panel demographics, as predetermined by the organization is small enough, then the NPS conducted using the same design as on the first occasion, provides valid statistical inferences without the need for a PS. In the next sections we describe the methodology, then make some comparisons using health data from the Los Angeles MSA and conclude with suggestions for future research.

1.1 The Proposed Methodology

Using data from a NPS for statistical inference can involve several levels of use including using the NPS survey data to L1) make overall estimates for the population of interest, L2) making subpopulation estimates and L3) using the NPS data set for multivariate analysis. We assume that the NPS is a quota sample from a panel where the overall demographics of the panel for the target population of interest are known. This rules out the use of river sampling for the NPS. We also assume that there is a contemporaneous PS, the gold standard that can be used for comparison purposes. The organization can then select variables for comparison for levels L1 and L2 and also decide on what comparisons will be made for L3 if chosen for use. For each level an *a priori* decision rule is developed based on the selected variables. The rules are indexes created as follows: for each "bad" comparison the index is increased by some pre-determined value, say 1. For each "good" comparison there is no increment to the index.

We illustrate the methodology with some health variables, specifically the following eight variables: ever diagnosed with asthma, diabetes or cancer, ever smoked, current smoker, obesity, state of current health fair/poor health and visited doctor in past year[2]. For level L1 we compare 95% confidence intervals from the two surveys, denoted by an index, say I_{L1}. If confidence intervals do not overlap a 1 is added to L_1, if they do overlap a 0 is added.

[1] And not a river sample.

[2] These variables, or at least harmonized variables, should also be in the gold standard PS.

Since there are eight variables the maximum value of index I_{L1} is eight. The next decision to be made is to decide on what level of risk one is willing to accept to use the NPS for overall estimates: a low index score, say $I_{L1} \leq 2$, indicates low tolerance of risk, a higher index score, say, $I_{L1} \leq 4$, indicates a higher tolerance of risk.

For subpopulation estimates a similar index is created, say I_{L2}. The index is based on the subpopulations of interest. Suppose in this case the subpopulation of interest is estimates of the eight health variables by gender. The NPS and PS comparisons are based on 95% confidence intervals, an overall adds 0 to the index, and a non-overlap adds 1. Then the maximum value of I_{L2} is 16. As an example the cut off, the risk, for accepting the gender estimates, can be set at $I_{L2} \leq 4$.

Suppose the organization also wants to use the NPS data for multivariate analysis. An index can be created based on comparing correlations in the two data sets or can be developed directly based on the specific multivariate analysis needed. In this case the interest is understanding the demographic drivers of the health variables where the independent variable are gender (male vs. female), age (18 – 44 vs 45 +), education level (less than some college vs. some college or higher), Hispanic origin and Non-Hispanic Blacks. We illustrate the approach using ever diagnosed with diabetes and choose to compare the 95% confidence intervals of the odds ratios of the independent variables. We create an index, I_{L3}.where a 1 is added to the index of the confidence intervals do not overlap, otherwise nothing is added. The maximum value of I_{L3} for the diabetes model is 5.

Obviously an organization that wants to use this data now has a choice to make, do they want to evaluate each of the levels separately, so that the NPS can be used for one, two or three levels depending on the results from each level or do they want to use the NPS for all three levels? For the latter situation they will evaluate a summary index $I_L = I_{L1} + I_{L2} + I_{L3}$. Suppose that the overall level of risk the organization is willing to take is $I_L \leq 5$. If this occurs than the NPS data would be used for all three levels. In the next section we make the actual comparisons between the NPS and the PS for each of the three levels

1.2 Comparison of the NPS Data with the PS Data

In this section we compare the results from the two surveys. Table 1 compares the overall prevalence rates for the eight variables. Two variables, health status and diagnosed with cancer are significant so the index score is $I_{L1} = 2$.

Table 1: Comparison of 95% Confidence Intervals for Eight Health Variables and the Level 1 Index Score

Asthma	*Estimate*	*Lower Bound 95% C. I.*	*Upper Bound 95% C.I.*	*Index Scoring*
Probability Survey	7.2	6.0	8.3	
Non-Probability Survey	5.8	3.8	7.9	0.0
Diabetes				
Probability Survey	10.6	9.2	12.0	
Non-Probability Survey	12.1	8.6	15.6	0.0

Current Smoker				
Probability Survey	11.8	10.3	13.4	
Non-Probability Survey	12.1	8.6	15.6	0.0
Ever Smoker				
Probability Survey	34.2	32.0	36.4	
Non-Probability Survey	34.6	30.0	39.1	0.0
Fair/Poor Health				
Probability Survey	21.6	19.6	23.5	
Non-Probability Survey	14.9	10.5	19.3	1.0
MD Visit				
Probability Survey	66.0	63.7	68.2	
Non-Probability Survey	69.2	64.6	73.8	0.0
Obese				
Probability Survey	25.0	22.9	27.0	
Non-Probability Survey	29.3	24.7	33.9	0.0
Cancer				
Probability Survey	4.4	3.4	5.5	
Non-Probability Survey	7.9	5.9	9.9	1.0

The next set of comparisons is for level L_2, the subpopulation comparisons. Figures 1 and 2 show the comparisons of the eight health variables by gender, Figure 1 for women and Figure 2 for men. Of the 16 comparisons 1 significantly different, women diagnosed with cancer. The L2 index score then is $L_2 = 1$.

The final set of comparisons, for level L3, is based on a logistic regression where for purposes of illustration the dependent variable is ever diagnosed for diabetes and the independent variables are gender (male vs. female), age (18 – 44 vs 45 +),

Figure 1: Comparison of Health Variables for Women: Estimates and 95% Confidence Intervals

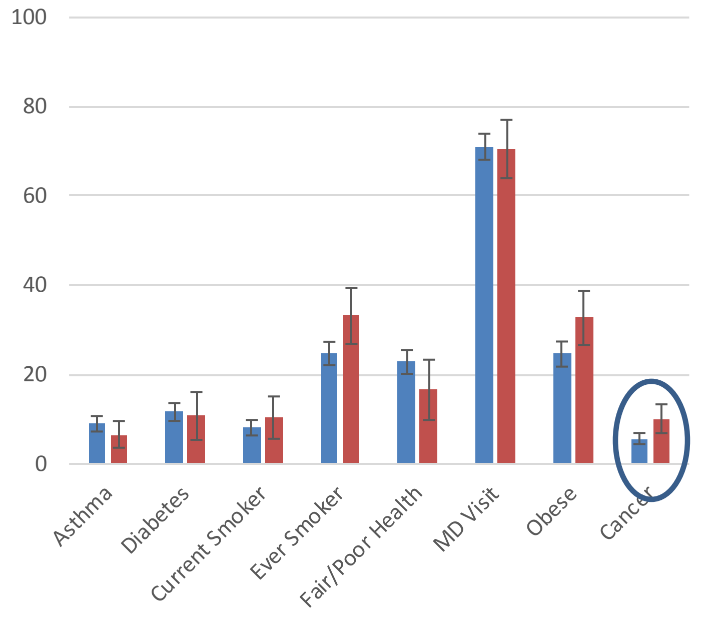
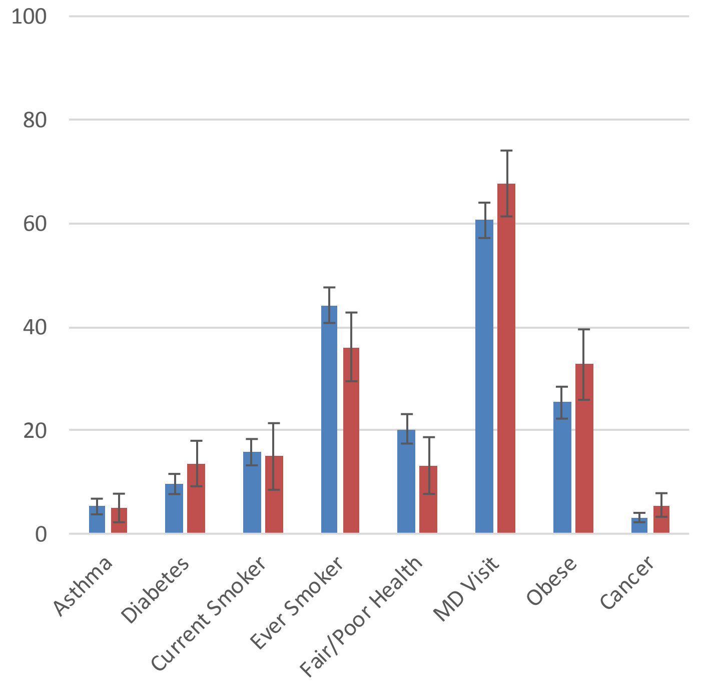


Figure 2: Comparison of Health Variables for Men: Estimates and 95% Confidence Intervals



education level (less than some college vs. some college or higher), Hispanic origin and Non-Hispanic Blacks. Table 2 shows the comparisons of the odds ratios of the logistic regression. Hispanic Origin is significant for the PS and not significant for the NPS, all other independent variables agree with respect to significance so for L3 L_3 = 1 out of a maximum score of 5.

Table 2: Comparison of 95% Confidence Intervals of the Odds Ratios for Dependent Variable Ever Diagnosed with Diabetes

	PS Diabetes		
	Odds Ratio	Lower 95% CI	Upper 95% CI
Male	0.791	0.578	1 .082
< College	1.678	1.206	2.333
Hispanic Origin	1.88	1.335	2.648
Black Non-Hispanic	1.543	0.978	2.434
18 – 44 Years	0.144	0.093	0.223
	NPS Diabetes		
Male	1.749	0.934	3.273
< College	1.693	1.03	2.781
Hispanic Origin	2.139	0.999	4.579
Black Non-Hispanic	1.087	0.482	2.451
18 – 44 Years	0.097	0.042	0.222

For the three levels combined the overall index score $I_L = I_{L1} + I_{L2} + I_{L3}= 2 + 1 + 1 = 3$. So if the overall level of risk for I_L was chosen, *a priori*, less than 3 the NPS is valid for all three levels of use. In the next section we outline a decision rule that would allow the NPS to be used on later occasions without the need for a gold standard PS.

1.3 Using the NPS on Later Occasions

The final step in the proposed empirical methodology is deciding on an *a priori* rule for using the NPS for inference without the need for a gold standard PS for comparison. This assumes that the comparison with the gold standard was successful. The rule proposed here is based on a comparison of the change in the demographic variables for the panel between the first comparison of the NPS and the PS and the second use of the NPS without the gold standard. An index is again created based on the 95% confidence intervals for a selected set of variables, say those used in the logistic regression in section 1.2 or some larger set. Again we create an index based on comparing demographics for the target area, if the intervals overlap nothing is added to the index, if they do not overlap a positive score is added and *a priori* we set a cut-off level where a score less than or equal to the cut-off indicates that the panel has not changed enough to prohibit its use. Going one step further the NPS should be conducted in the same manner as is was conducted when it was compared to the PS. The NPS on this occasion now stands alone for inference. This process can be repeated on later occasion and as long as the cut-off score for the demographic comparisons are small enough we can continue conducting the NPS without the need for a PS comparison. Once the demographic comparison scoring exceeds the cut off score than we need to conduct another gold standard PS side by side with a NPS. The next section briefly summarizes the methodology and provides some recommendations for future research.

2. Summary and Future Research

2.1 Summary

We have proposed an empirical methodology based on *a priori* decision rules that moves the research on NPS surveys from the comparison stage to the use of the data for statistical inference. At the first stage comparisons can be made for estimates for the target population and subpopulations and for multivariate analysis. The NPS should be a quota sample from a panel and its results are compared to a contemporaneous PS. If this first level of comparisons are positive then a comparison of the demographics of the panel at a later time, with little or no change, indicates that the NPS can again be used for inference without the need for a gold standard PS, that is, the NPS is of sufficient quality to stand alone.

2.1 Future Research

There are several areas of future research that should be addressed. The first is that a potential major effect of mode should be eliminated if possible when comparing the NPS to the PS. Since the most NPS are web-based it would enhance the comparison of the two surveys if the PS was conducted in the same mode: an Addressed Based Sample (ABS)using a push to web or an ABS self-administered mail questionnaire designed to minimize the stimulus presented to respondents in both surveys.

The decision rules can also be expanded by adding comparisons of the ratios of response or contact rates between the two surveys or by comparing the post-stratification weighting adjustments by using the ratio of the coefficients of variation of the weights as a part of the *a priori* decision rules.

In addition there might be scoring methods that could be considered beyond comparing confidence intervals. A rule like assigning points based on if NPS estimate falls within PS 95% CI and also how close the NPS estimate is to PS estimate such as assign a 0 if NPS within 95% CI of PS, a 1 if NPS outside 95% CI but within 10%(absolute) or 3% (relative) points and a 2 if NPS outside 95% CI and differs by >10% / 3% points.

Finally, when using an index that combines more than one level of use of the NPS data it may be useful to assign a relative importance to each level rather than using an overall index $I_L = I_{L1} + I_{L2} + I_{L3}$ giving equal weight to each level it may be useful to weight each level.

Acknowledgements

Robert Tortora would like to personally thank Joe Sedransk for helpful discussions early in the development of the methodology.

.

References

Baker, R, D. Zahs, and G. Papa, 2004. “Health Surveys in the 21st Century: Telephone vs. Web.” Paper presented at the Eighth Conference on Health Survey Research Methods, Peachtree City, GA. Accessed at: http://www.cdc.gov/nchs/data/hsrmc/hsrmc_8th_proceedings_2004.pdf. (pg. 143-148)

Bergdahl, H., P. Biemer, and D. Trewin, 2014. “Pushing Forward with ASPIRE” Accessed at http://www.q2014.at/fileadmin/user_upload/Pushing_forward_with_ASPIRE_Q2014.pdf)

Berrens, R., Alok K. Bohara,A. K., Jenkins-Smith, H., Silva, C. and Weimer, D.L., 2003. "The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Samples." *Political Analysis*. 11.1 (2003): 1-22.

Bethlehem, J. and Biffignandi, S., 2010. "Selection Bias in Web Surveys." *International Statistical Review*. 78.2: 161-188.

Chang, L. and Jon A. Krosnick, J.A., 2009. "National Surveys via RDD Telephone Interviewing versus the Internet: Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly*. 73.4: 641-678.

Couper, M. P., 2007. "Issues of Representation in eHealth Research (with a Focus on Web Surveys)." *American Journal of Preventive Medicine*. 32.5. (May): S83-S89.

Duffy, B., Smith, K., Terhanian, G. and Bremer, J., 2005. "Comparing Data from Online and Face-to-Face Surveys." *International Journal of Market Research*. 47.6: 47-62.

Elliott, M. R., 2009. "Combining Data from Probability and Non-Probability Samples Using Pseudo-Weights." *Survey Practice*. 2.6: 1-7.

Iachan, R., Boyle, J. and Harding. L., (2016). "Inferences from Internet Panel Studies and Comparisons with Probability Samples". Presented at the Joint Statistical Meetings, Chicago, IL.

Lee, S., 2006. "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys." *Journal of Official Statistics*. 22.2: 329-349.

Lee, S. and Valliant, R., 2009. "Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment." *Sociological Methods Research*. 37.3: 319-343.

Loosveldt, G. and Sonck, N., 2008. "An Evaluation of the Weighting Procedures for an Online Access Panel Survey." *Survey Research Methods*. 2.2: 93 – 105.

Stephan, F. and McCarthy, P., 1958. Sampling Opinions, Chapter 8. Wiley Series in Statistics, NY.

Sudman, S., (1965). Probability Sampling with Quotas. JASA, Vol. 61, 319, 749 – 771.

Smith, T.M.F., 1983. "On the Validity of Inferences from Non-Random Sample." *Journal of the Royal Statistical Society*. Series A, 146.4: 394-403.

Valliant, R., and Dever, J. (2011), "Estimating Propensity Adjustments for Volunteer Web Surveys," Sociological Methods and Research, 40, 105-137. Accessed at https://jpsm.umd.edu/publication/valliant-r-and-dever-j-2011estimating-propensity-adjustments-volunteer-web-surveys