Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

Estimation with Non-probability Surveys and the Question of External Validity

Jill A. Dever and Richard Valliant[1]

Abstract

Probability-based surveys, those including with samples selected through a known randomization mechanism, are considered by many to be the gold standard in contrast to non-probability samples. Probability sampling theory was first developed in the early 1930's and continues today to justify the estimation of population values from these data. Conversely, studies using non-probability samples have gained attention in recent years but they are not new. Touted as cheaper, faster (even better) than probability designs, these surveys capture participants through various "on the ground" methods (e.g., opt-in web survey). But, which type of survey is better?

This paper is the first in a series on the quest for a quality framework under which all surveys, probability- and non-probability-based, may be measured on a more equal footing. First, we highlight a few frameworks currently in use, noting that "better" is almost always relative to a survey's fit for purpose. Next, we focus on the question of validity, particularly external validity when population estimates are desired. Estimation techniques used to date for non-probability surveys are reviewed, along with a few comparative studies of these estimates against those from a probability-based sample. Finally, the next research steps in the quest are described, followed by a few parting comments.

Key Words: Non-probability sample surveys; Fit for purpose; Quality framework; External validity.

1. Introduction

Innovation! This word is heard and used constantly in today's world. Many books are written on the subject, posing the question: how we can make things better? For example, in a recent book by Bryce (2014), innovation (and hence the definition of "better") is linked to the characteristics, "Smaller Faster Lighter Denser Cheaper." This same sentiment holds for survey research. However, how might we define "better" in the survey context?

Holt (2007), for example, defines "better" using terms such as wider, deeper, quicker and cheaper. He sees positive benefits of "relentless pressures" placed on government agencies to produce high quality official statistics in quick succession as one way to increase innovation for survey research. This opinion was further emphasized by Constance F. Citro, director of the Committee on National Statistics[2], in her Waksberg Award Winner Address at the 2014 Statistics Canada Symposium. Such pressures are not just relegated to those producing official statistics but they are felt by the entire research community.

Policymakers, health professional, government agencies, and the public at large all require rapid information. For example, the U.S. Centers for Disease Control conducts a general population survey on barriers to influenza vaccinations early in the season to inform public health educational campaigns each December (Srivastav et al., 2014). Rapid information is no more critical than when natural disasters or infectious diseases hit the news airways. Most recently, surveys were released in the U.S. to assess the public's understanding about the likelihood of contracting the Ebola virus (see, e.g., Dwyer, 2014) and healthcare professional preparedness to handle suspected cases (see, e.g., Garman, 2014). However, decreasing participation and ever growing resource needs (cost and time) associated with probability samples cause some to question the utility of the more traditional designs.

[1]Jill A. Dever, RTI International, 701 13th St NW, Suite 750, Washington, DC 20005-3967, jdever@rti.org; Richard Valliant, Institute for Social Research at the University of Michigan, Joint Program in Survey Methodology at the University of Maryland, 1218 LeFrak Hall, College Park, MD 20742, rvallian@umd.edu.
[2] http://sites.nationalacademies.org/dbasse/cnstat/index.htm

One answer researchers have given to this call for innovation is the non-probability sample survey. The use of non-probability samples has a long history, as noted in Smith (1976). Touted as cheaper, faster (even better) than studies with probability-based samples, these surveys capture participants through various "on the ground" methods (e.g., opt-in web survey). One extreme case of a convenience sample that predicted the outcome of the 2012 U.S. presidential election well was of Xbox gaming console users (Wang, et al. 2015). Favorable results such as this encourage some people to believe that probability sampling could be jettisoned entirely. But, are we trading quality or generalizability (i.e., external validity) in this quest for quicker, cheaper data? We contend that the answer to this question is complicated ("it depends" if you will). Or, as the Statistics Canada (2009) Quality Guidelines states "quality is relative, not absolute."

The answer depends on a variety of existing quality measures reviewed in the first section of our paper. Next, we discuss two important dimensions within the fit-for-purpose paradigm and note some complicating factors for non-probability surveys (section 3). Techniques currently in use to produce estimates from non-probability surveys are summarized in section 4 to frame the issues. Where relevant, mixed results from these methods when compared against their probability-based counterpart are noted (section 5). We highlight a few questions that need to be addressed in future research (section 6) and conclude by summarizing our thoughts.

2. Survey Quality

Art for some people cannot be easily defined, but they know it when they see it. In the past, the same could be said for survey quality. Some researchers initially worked to identify one overarching quality measure that was easily comparable across surveys. But, such an endeavor has proven difficult. Take, for example, the response rate. Groves & Peytcheva (2008) demonstrated that the response rate is not necessarily a good surrogate measure for data quality because there is not always a strong correlation between a low rate and the increased presence of nonresponse bias in the data.

Statistics Canada (2009) Quality Guidelines notes, "During the past twenty years, statistical agencies have arrived at a consensus that the concept of *quality* of statistical information is multi-dimensional" (see also Biemer and Lyberg, 2003). However, not all quality dimensions are created equal. Enter the phrase – fit(ness) for purpose – into the survey researcher's vocabulary for the development of study designs (Baker et al., 2013). This phrase has existed for a long time to describe quality (or lack thereof) in a variety of areas. Fit for purpose means that something is "good enough to do the job it was designed to do" (http://www.macmillandictionaryblog.com/) and "appropriate, and of a necessary standard, for its intended use" (http://en.wiktionary.org/). The fit for purpose approach identifies (or at least should) which in a series of quality dimensions is most important. For example, a survey related to an infectious disease outbreak may prioritize real-time estimation over other dimensions.

In the next two sections, we review two quality frameworks from the literature to define the dimensions of fit for purpose. To date, these frameworks are primarily used for probability-based surveys. The last section briefly highlights some challenges with defining a comparable framework for non-probability surveys.

2.1 Statistics Canada Quality Assurance Framework

Statistics Canada (2002) Quality Assurance Framework cites six interconnected measures of quality:

- Relevance – data meets the needs specified by the client (e.g., funding agency, research community, and the public).
- Accuracy – data can be used to correctly measure target population as intended with acceptable levels of precision and accuracy.
- Timeliness – data pertain to time period in question and are available when needed.
- Accessibility – data easily available using a suitable data collection mechanism with reasonable cost.

- Interpretability – data/survey design is easily understood.
- Coherence – consistency with other data sources.

2.2 Total Survey Error Quality Framework

Total survey error (TSE) encompasses the body of research oriented toward detection and treatment of a host of errors that can occur with surveys including the sampling errors and challenges associated with nonresponse. Biemer (2010) discusses a quality framework through the TSE lens that includes the six dimensions identified by Statistics Canada (section 2.1), along with three others:

- Credibility – data are considered to be reliable.
- Comparability – methodologies are consistent across similar studies.
- Completeness – data are available to meet analytic objectives without "undue burden on respondents."

2.3 Quality Frameworks for Non-probability Surveys

Baker et al. (2013), for example, discuss a few challenges for assessing the quality of non-probability surveys. For example, with many non-probability surveys the sampling frame is non-existent and therefore not available for scrutiny by the research community. We provide additional thoughts below on a few dimensions listed previously:

- Accuracy/comparability – estimates from non-probability surveys have sometimes matched comparable estimates from other sources and sometimes not (see section 4 for this discussion).
- Coherence – challenges with explaining inconsistencies with other data sources is an issue for both probability and non-probability surveys.
- Credibility – the reliability for surveys in general come in to question if the validity of the data is not justified or if the level of nonparticipation (if known) is deemed excessive by the research community.

Words such as timeliness, accessibility, and completeness are sometimes used to justify the choice of a non-probability survey over one with a defined random sampling mechanism.

3. Generalization of Survey Estimates

More and more, study designs (i.e., a combination of the essential survey conditions like the sampling design, data collection mode, and the like [Hansen et al., 1961]) are chosen to maximize a survey's fit for purpose. Two quality dimensions important to fit for purpose and of focus for this paper are validity and accuracy.

3.1 Validity

Validity comes in two flavors and has a direct effect on the choice of the sampling design. The phrase *internally valid* (see, e.g., http://www.edpsycinteractive.org/) suggests that all confounders were measured (and measured well) on your participants and are available for analyses. This phrase can be found throughout study literature on, for example, clinical trials. The label *externally valid* (see, e.g., http://www.socialresearchmethods.net/) indicates that the study estimates are generalizable outside the sample, say to the population of interest, and in fact are reproducible.

Researchers using non-probability surveys are especially subject to criticism on the lack of external validity. One reason is that many of these studies do not have a sampling frame and hence lack a discernable link to the target population (see, e.g., Baker et al., 2013). Participants in the non-probability surveys may have been "captured" by coincidence and these same participants might not be accessible if the survey were administered again. One link

from a non-probability sample to the population is the hope that a common model can be used to predict the values of analysis variables for both sets of units (e.g., see Valliant, et al. 2000). However, diagnosing whether a model(s) holds for both may be difficult or impossible. Hence, defense of the "reproducibility assumption" may be weak.

3.2 Accuracy

Accuracy, a function of bias (squared) and precision, is the difference between the estimate and the (true) intended parameter of interest. In the survey world, an estimate is considered to be accurate if the mean square error (=bias^2 + variance) is small. We briefly discuss precision for non-probability surveys in the next section and leave a more in-depth examination for another day. This leaves bias.

To describe three components of survey bias, Valliant and Dever (2011) first define three populations: (1) the *target population* of interest for the study; (2) the *potential covered population* available by way of the essential survey conditions; and (3) the *actual covered population*, the portion of the target population that is recruited for the study through the essential survey conditions. For example, consider an opt-in web survey for a smoking cessation study. The target population may be defined as adults aged 18-29 who currently use cigarettes. The potential covered population would be those study-eligible individuals with internet access who visit the sites where study recruitment occurs; those actually covered would be the subset of the potential covered population who participate in the study.

Having specified the three populations, the authors then define three bias components as:

- Coverage bias – difference between the target population and the potential covered population.
- Nonresponse / nonparticipation bias – difference between the potential and actual covered populations.
- Selection bias – difference between the target population and actual covered population.

The survey literature is full of studies to assess the level of coverage bias associated with certain sampling frames such as landline telephone numbers in the U.S. (see, e.g., Christian et al., 2010) and with certain data collection modes such as the internet (see, e.g., Dever et al., 2008). Many citations exist for adjusting probability-based weights for nonresponse bias (see, e.g., Kott, 2006). The term selection bias is found throughout the non-probability survey literature (see, e.g., Lee and Valliant, 2009) since for many studies characteristics of the nonparticipating (potential) sample members or even the percent of nonrespondents is unknown.

Many researchers using surveys with probability-based samples rely on design- or model-assisted theory and/or models to limit errors in the data to justify the design unbiasedness of the estimates (see, e.g., Särndal et al., 1992). For example, weight calibration can be used to minimize bias associated nonresponse and coverage (Kott, 2006). In the next section, we summarize adjustment techniques for non-probability surveys that at least attempt to correct for selection bias, along with a few studies that evaluate the utility of such procedures.

4. Estimation Methods for Non-probability Surveys

Below we summarize some of the adjustment and estimation methods used to date to maximize the utility of data obtained from non-probability surveys. For convenience, we classify these methods as either model-based or pseudo design-based techniques.

4.1 Model-based Methods

The first method is referred to as the *population-model approach*. Here, researchers use the survey data with in an established model for the population of interest. The assumptions needed for this approach, as discussed in many mathematical statistics textbooks, include a random sample from the population (i.e., data in the sample and the population follow the same model, which is true if data not included in the sample are missing at random [see, e.g., Valliant et al., 2013]), and any possible confounders are measured and included in the model. Though population estimates are generated, verification of the assumptions may not be discussed. Relevant examples of the population-

model approach are too numerous to name in this short recitation. The theory for this approach is well established (Valliant et al. 2000); methods for diagnosing whether a model for a non-probability sample holds for the full population are not.

Another model-based method is known as a *Heckman model* (1979). Developed for economic modeling with observational data, the mechanism for inclusion in the study is estimated using an initial model. The resulting correction for selection bias is then included in a subsequent estimation model. This technique heavily relies on the assumption of normality in the underlying distribution to estimate selection bias. As with the first method, the assumption of randomly missing records from the sample holds here as well.

The last model-based approach we briefly comment on is *Bayesian modeling*. As with traditional Bayesian modeling or superpopulation modeling, the units in the population and the sample must follow the same model. Priors on model parameters can lead to richer classes of estimators than if no priors are used. To date this approach is not widely used but is gaining interest (see, e.g., Little, 2004; Roshwalb et al., 2012).

Before we conclude our brief discussion of model-based methods for non-probability surveys, we must address the issue of missing at random. Some surveys with undefinable randomization mechanisms come with claims that the non-participants are missing at random because the demographic distributions align with the population of interest. However, as discussed in Dever et al., (2008), demographic information may not be an effective litmus test (or weight adjustment) for the needed missingness. Tests for not missing at random do exist but currently only for surveys with a random sample (Pfeffermann and Sikov, 2011).

4.2 Pseudo Design-based Estimation

The missing at random assumption is also used within what we call pseudo design-based estimation. Here, researchers use data from a survey without a "defined" random sampling mechanism and treat them as if there were collected from a *(stratified) simple random sample* (see, e.g., Thompson, 2002). In other words, the non-probability survey is reclassified as having a simplistic single-stage sampling design when in many cases there was no frame from which to draw such a sample. The base weights, generally set to a value of one, are calibrated to population totals (Loosveldt and Sonck 2008). The calibrated analysis weights are then used with the data and traditional Horvitz-Thompson estimators (1952) to calculate population statistics (Deville and Särndal, 1992). This method can also be considered as model-based in the sense that the calibrating variables imply a model that must hold for both the sample and nonsample parts of the population.

The next technique is known as the *propensity score adjustment* method and is based on the work by Rosenbaum and Rubin (1983) for observational studies. Adapted for survey research, a non-probability sample is combined with a (reference) probability sample to estimate the propensity of being a non-probability survey respondent. The inverse of the propensity is used either directly to form the analysis weight or indirectly to form poststrata (see, e.g., Lee & Valliant 2009). Again, the analysis weight may be calibrated to population totals and the calibrated analysis weight used to create Horvitz-Thompson like estimates.

5. Comparative Studies

Historically, the literature to document the utility of adjustments for non-probability sample surveys is sparse but the tide is turning. For example, Tourangeau et al. (2013) examined eight opt-in (volunteer) web panels using weight adjustment techniques and report that the results were mixed. In general, adjustments removed only part of the selection bias and in some cases actually increased the bias relative to the unadjusted estimates. Benefits of the adjustment also varied widely by type of analysis variable. Yeager et al. (2011) had similar mixed results for the comparison of an RDD survey with a non-probability internet survey.

Valliant and Dever (2011) summarize the assumptions required to adapt the Rosenbaum and Rubin (1983) methodology to the world of surveys for the propensity score adjustment method. Namely, both surveys need to include randomly chosen participants where nonparticipants are missing at random, and samples for both surveys

cannot overlap. The latter is noted as a clear violation for most if not all non-probability-reference survey situation. The authors noted that the analysis weights for the reference survey were required in the estimation of the propensity scores to minimize selection bias. Research is still needed to determine the best input weight for the non-probability sample for propensity score methods.

6. Future Research

Other comparative studies exist in the literature suggesting mixed results for the external validity of non-probability-based estimates both within and across the surveys assessed. Future research may include the daunting task of determining when non-probability estimates work well. Valliant (2013) suggests that models may be an important tool in answering this mystery. Also, the answer must include a quality framework for all surveys designed toward the fit for purpose goals.

Notably, survey researchers interested in defining the appropriate place (if one exists) for sample surveys without a defined random sampling mechanism (i.e., without a probability-based sample) have much work to do. We define for goals for future research in terms of a series of questions, some of which we hope to add to the knowledge base soon.

(1) Do quality dimensions exist to appropriately compare probability and non-probability surveys, and non-probability surveys against each other?

(2) Are there criteria for relabeling a probability survey as non-probability based on, e.g., low response rates, and hence switching to this new quality framework?

(3) Within the quality framework, how might we define a set of standard criteria for "accepting" statistics from a non-probability survey as population estimates?

Of potential relevance to (3) are balance measures which have been proposed for the study of nonresponse (Särndal and Lundquist 2014) and have a key role in optimality in the model-based approach (Royall 1992).

7. Conclusion

Innovation has brought survey research a long way and will continue to do so in the future. Non-probability surveys are seen by some as an innovative way to address growing pressures for cheaper, faster data especially as nonresponse to probability surveys is ever increasing. Many questions and much debate exist, however, on whether non-probability surveys are just a quick fix and not a real solution. This article is the first in a series to define solutions and an associated quality framework for all surveys in terms of a study's fit for purpose. Here, we summarize the mixed results obtained from research on non-probability surveys to date to frame the discussion. Much work is still needed, but have we ever backed down from a challenge?

References

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., et al. (2013), "Summary report of the AAPOR task force on non-probability sampling", *Journal of Survey Statistics and Methodology*, 1, pp. 90–143.

Biemer, P. P. (2010), "Total Survey Error: Design, Implementation, and Evaluation", *Public Opinion Quarterly*, 74(5), pp. 817–848.

Biemer, P. P. and Lyberg, L. E. (2003), *Introduction to Survey Quality*, New Jersey: John Wiley & Sons, Inc.

Bryce, R. (2014), *Smaller Faster Lighter Denser Cheaper: How Innovation Keeps Proving the Catastrophists Wrong*, New York: PublicAffairs™.

Christian, L., Keeter, S., Purcell, K. and Smith, A. (2010), "Assessing the Cell Phone Challenge", Pew Research Center report available at http://www.pewresearch.org/2010/05/20/assessing-the-cell-phone-challenge/.

Dever, J.A., Rafferty, A., and Valliant, R. (2008), "Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?" *Survey Research Methods*, 2(2), pp. 47-62.

Deville, J.-C. and Särndal, C.-E. (1992), "Calibration estimators in survey sampling", *Journal of the American Statistical Association*, 87, pp. 376-382.

Dwyer, M. (2014), "Poll finds many in U.S. lack knowledge about Ebola and its transmission", Harvard School of Public Health press release (August 21, 2014) available at http://www.hsph.harvard.edu/news/press-releases/poll-finds-many-in-us-lack-knowledge-about-ebola/.

Garman, L. (2014), "APIC [Association for Professionals in Infection Control and Epidemiology] Ebola Readiness Poll: Results of an online poll of infection preventionists", Available at http://www.apic.org/Resource_/TinyMceFileManager/Topic-specific/Ebola_Readiness_Poll_Results_FINAL.pdf.

Groves, R. and Peytcheva, E. (2008), "The impact of nonresponse rates on nonresponse bias: A meta-analysis", *Public Opinion Quarterly*, 72 (2), pp. 167–189.

Hansen, M.H., Hurwitz, W.N., and Bershad, M.A. (1961), "Measurement Errors in Censuses and Surveys", *Bulletin of the International Statistical Institute*, 38(2), pp. 359-374.

Heckman, J.J. (1979), "Sample Selection Bias as a Specification Error", *Econometrica*, 47, pp. 153–62.

Horvitz, D.G. and Thompson, D.J. (1952), "A generalization of sampling without replacement from a finite universe", *Journal of the American Statistical Association*, 47, pp. 663–685.

Holt, D.T. (2007), "The Official Statistics Olympic Challenge: Wider, Deeper, Quicker, Better, Cheaper", *The American Statistician*, 61(1), pp. 1-8.

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

Lee, S. and Valliant, R. (2009), "Estimation for volunteer panel Web surveys using propensity score adjustment and calibration adjustment", *Sociological Methods & Research*, 37(3), pp. 319-343.

Little, R.J.A. (2004). "To Model or Not Model? Competing Modes of Inference for Finite Population Sampling", *Journal of the American Statistical Association*, 99, 546-556.

Loosveldt, G. and Sonck, N. (2008), "An evaluation of the weighting procedures for an online access panel survey", *Survey Research Methods*, 2(2), pp. 93-105.

Pfeffermann, D. and Sikov, A. (2011), "Imputation and Estimation under Nonignorable Nonresponse in Household Surveys with Missing Covariate Information", *Journal of Official Statistics*, 27(2), pp. 181–209

Rosenbaum, P.R. and Rubin, D.B. (1984), "Reducing bias in observational studies using subclassification on the propensity score", *Journal of the American Statistical Association*, 79, pp. 516-524.

Roshwalb, A., El-Dash, N., and Young, C. (2012), "Towards the Use of Bayesian Credibility Intervals in Online Survey Results", white paper produced by Ipsos Public Affairs and available at http://www.ipsos-na.com/dl/pdf/knowledge-ideas/public-affairs/IpsosPA_POV_BayesianCredibilityIntervals.pdf.

Royall, R.M. (1992), "Robustness and Optimal Design under Prediction Models for Finite Populations", *Survey Methodology*, 18, 179-185.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag, Inc.

Särndal, C.-E., and Lundquist, P. (2014), "Accuracy in Estimation with Nonresponse: A Function of Degree of Imbalance and Degree of Explanation", *Journal of Survey Statistics and Methodology*, 2, 361-387.

Smith, T.M.F. (1976), "The Foundations of Survey Sampling: a Review". *Journal of the Royal Statistical Society A*, 139, 183-195.

Srivastav, A., Santibanez, T.A., Kahn, K.E., Zhai, Y., Greby, S.M., et al (2014), "National Early Season Flu Vaccination Coverage, United States, November 2014". Available at http://www.cdc.gov/flu/pdf/fluvaxview/nifs-estimates-nov2014.pdf.

Statistics Canada (2002), "Statistics Canada's Quality Assurance Framework", available at http://www5.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-586-X&CHROPG=1&lang=eng.

Statistics Canada (2009), "Statistics Canada Quality Guidelines", available at http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.htm.

Thompson, S.K. (2002), *Sampling*, New York: John Wiley & Sons, Inc.

Tourangeau, R., Conrad, F.G. and Couper M.P. (2013), *The Science of Web Surveys*, New York: Oxford University Press.

Valliant, R. (2013), "Summary report of the AAPOR task force on non-probability sampling (Comment)", *Journal of Survey Statistics and Methodology*, 1, pp. 105–111.

Valliant, R., & Dever, J.A. (2011), "Estimating propensity adjustments for volunteer web surveys", *Sociological Methods & Research*, 40(1), pp. 105–137.

Valliant, R., Dever, J. A., & Kreuter, F. (2013), *Practical tools for designing and weighting sample surveys*, New York: Springer.

Valliant, R. Dorfman, A., and Royall, R.M. (2000), *Finite Population Sampling and Inference: A Prediction Approach*, New York: John Wiley.

Yeager, D.S., Krosnick, J.A., Chang, L., Javitz, H.S., Levendusky, M.S., Simpser, A., & Wang, R. (2011), "Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples", *Public Opinion Quarterly*, 75(4), pp. 709–747. Available at http://www.knowledgenetworks.com/insights/docs/Mode-04_2.pdf.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015), "Forecasting Elections with Non-representative Polls", *International Journal of Forecasting*, forthcoming.