Enhancing Respondent Representativeness Through Responsive Design And External Benchmarks

Julia Shin-Jung Lee*

Abstract

Ongoing efforts to increase response rates and concerns about rising nonresponse rates are all grounded in issues related to nonresponse bias. Nonresponse bias is an error caused by patterns and mechanisms of nonresponse which distort the sample representativity of the initial sampling design and selection. The problem of nonresponse bias has been addressed by both improvements in data collection methods and refinements to post-survey adjustment. This paper argues that improving respondent representativeness with respect to the target population is a more effective strategy to reduce nonresponse bias. We describe a responsive design strategy that uses high quality external "benchmark" data during survey design and data collection to improve representativeness of the respondent sample, and consequently attenuates nonresponse bias while minimizing the post-survey corrective adjustment. This design strategy is particularly useful in the context of a multi-phase survey. After the completion of the first phase of data collection with a replicate sample, a propensity score model is fit using the first phase respondent data and the benchmark data to estimate the propensity of 'benchmark survey' membership. The ratio of the benchmark and the current survey propensity score density points to the proportion of imbalance between the benchmark and the new survey respondents, which is used subsequently to determine the optimal sample allocation of the second phase in such a way to restore balance between the observed sample and the benchmark. As this procedure is repeated, the respondent distribution moves toward to the benchmark population, increasing the resemblance of the two propensity distributions. Drawing on two large government surveys (NHIS and BRFSS), illustrations of this sampling strategy via simulation studies suggest that implementation of the proposed method can effectively enhance the respondent representativeness.

Key Words: respondent representativeness, responsive design, adaptive sampling, propensity score, NHIS, BRFSS

1. Introduction

Survey respondent representativeness has become a major focus in survey research, a focus that can mostly be attributed to the problem of steadily increasing nonresponse rate in survey data collections (Pew Research Report, 2012; De Leeuw and De Heer, 2002; Groves and Couper 1998). The increasing nonresponse rates are of particular interest because they call into question the validity of survey inferences and raise the potential for nonresponse bias. Nonresponse bias arises when the nonresponse rate is high and respondents and nonrespondents systematically differ on the characteristics of interest.

Two common strategies in reducing nonresponse bias are 1) increasing the survey response rate and 2) applying post-survey nonresponse weighting adjustments. Intuitively, simply increasing the response rate appears to be a natural way to try to minimize potential nonresponse bias; however, researchers have in fact demonstrated that this approach can exacerbate the problem especially if the increase is disproportionate with respect to the subgroups of the population (Groves, et al. 2008; Bootsma-van der Wiel, et al. 2002; Barclay, et al. 2002; Groves, et al. 2000). This finding raises the possibility that achieving

*PhD candidate, Program in Survey Methodology, 426 Thompson St, Ann Arbor, MI 48104

a representative respondent pool may be much more effective at reducing nonresponse bias than simply targeting a high response rate.

The traditional methods to reduce nonresponse bias involve applying weighting adjustments during survey estimation. These methods rest upon unverifiable missing at random (MAR) assumptions about non-response mechanism, may require complex modeling in the adjustment steps, and often result in reduced precision.

This paper describes a novel responsive design method to improve final survey respondent representativeness using a high quality survey or census benchmark for the target population. The term "responsive design", first introduced by Groves and Heeringa (2006), refers to a survey design strategy that is implemented in phases. At each sequential phase of a responsive design, sample design features and survey procedures are modified with the aim of minimizing cost and errors for the final survey product. A responsive design that aims to align respondent data distributions with the population during survey data collection provides an opportunity to minimize the differences between respondent and population distribution. Such a design strategy is particularly useful in the context of a multi-phase survey where the data collection can be organized into several discrete phases so that a skewed respondent pool can be corrected as more phases of data collection take place.

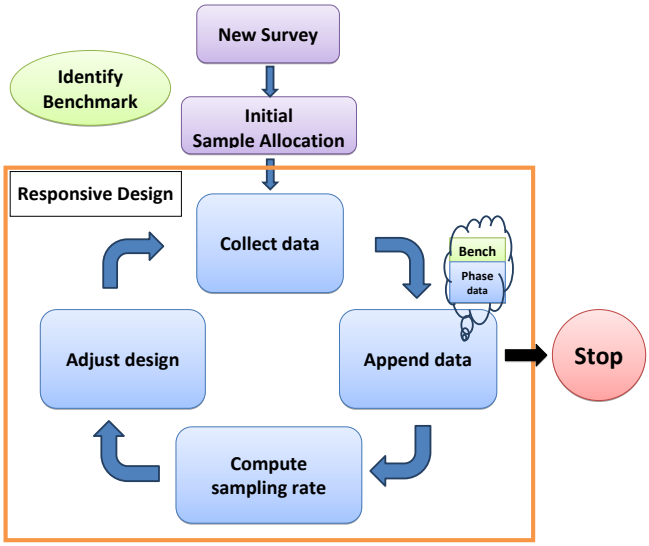

Figure 1: Proposed Responsive Design

The proposed responsive design starts with the first phase of data collection for the current survey based on an initial sample replicate. Using data from the first phase of the current survey and the benchmark survey, a propensity score model is fit to estimate the propensity of 'benchmark survey' membership. The ratio of the benchmark and the current survey propensity score density points to the proportion of imbalance between the surveys in terms of the target population characteristics. This ratio then serves as the basis of modifying the sampling rate, and adjusting the sample allocation in the second phase of data collection with the goal of improving the representativeness of the respondent pool. As this procedure is repeated, the respondent distribution moves closer to the benchmark population, providing a more representative respondent pool. The core process of the proposed responsive design is illustrated in the box in Figure 1.

In essence, the representativeness of the respondents is enhanced by a targeted proba-

bility sampling strategy using an external high quality survey or census that captures the target population as the benchmark. The multivariate nature of the target population is summarized in a single quantity through the propensity scores. The ratio of the propensity score density of the two surveys serves as a key factor of the sampling rate then guides sampling decisions to improve the representativeness of survey respondent at each data collection phase. In this study, the proposed sampling strategy is implemented hypothetically to mimic responsive design using two high quality government surveys, namely the National Health Interview Survey (NHIS) and the Behavior Risk Factor Surveillance Survey (BRFSS). The NHIS is used to estimate target population distributions whereas the BRFSS serves as the new survey.

Responsive design strategies targeting nonresponse bias reduction at the survey design stage have been investigated in several studies. These studies adopted very similar principles, including : 1) using auxiliary and contextual data available from both respondents and nonrespondents to estimate response propensity; 2) assigning a superior protocol to cases with low predicted response propensity in an effort to increase survey participation; and 3) altering design parameters to oversample subgroups of interest if necessary. The response rate is increased by case prioritization where cases with low predicted response propensity or low predicted contact probability receive different protocols. Examples of such studies include the National Survey of Family Growth Cycle 7 (NSFG C7), the Community Advantage Panel Survey (Peytchev, et al., 2010a), the Survey of Consumer Attitudes (SCA), and the newly redesigned 2013 National Survey of College Graduates (NSCG).

Reducing potential nonresponse bias by improving respondent representativeness is the primary objective of the responsive design implemented by some Norwegian surveys (for example, the Level of Living Survey, LLS). In these surveys, the respondent representativeness is evaluated by the *Representativeness* indicators, namely, R-indexes and partial R-indexes)(Schouten and Cobben, 2007). R-indicators estimate the similarity of the respondent and the sample pools with respect to auxiliary variables that are available from external sources for both respondents and nonrespondents. R-indicators are used as a monitoring device to identify characteristics of the underrepresented subjects and to optimize the field strategies. For example, R-indicators were computed after each contact attempt in a Dutch survey of Integrated Survey on Household Living Conditions (POLS). The plot of R-indexes against the response rate indicated a drop in representativeness for each additional contact attempt, which suggests a change in field work strategy such as a different data collection mode or incentive program. Similar to other responsive designs, the computation of R-indicators depends on the availability of rich auxiliary information for respondents and nonrespondents. Survey settings that can provide such data include face-to-face surveys, where interviewers may observe auxiliary information, regardless of response status, and surveys with comprehensive frame and registry data that are common in some European countries, such as the Scandinavian countries and the Netherlands.

Our proposed research also focuses on improving respondent representativeness. The major difference between our study and others can be visualized in Figure 2. Figure 2 illustrates data that are available in any survey, consisting of survey variables (Y), demographic and background variables from survey (X), and auxiliary and contextual data available from frame and external sources (Z). In figure 2, superscripts denote the data collection phases and subscripts indicate respondents (R) and nonrespondents (NR) data. Current responsive design studies, either those aimed at increasing response rate or at increasing respondent representativeness, exclusively focus on using data that are available from both respondents and nonrespondents, such as Z_R and Z_{NR}. For example, the R-index, which measures whether the respondents are a random sample of the sample pool, is the variance of estimated response propensity and is computed using Z_R and Z_{NR}.

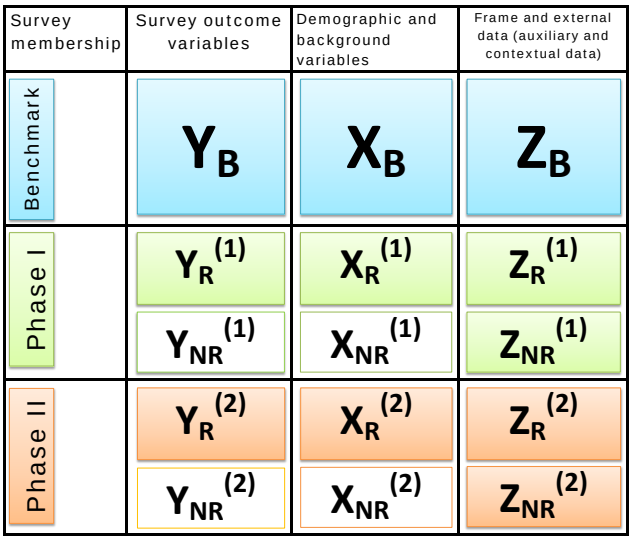

Figure 2: Data Available for Responsive Design

Our method improves the respondent representativeness through an external benchmark data source, such as Z_B. One advantage of using an external benchmark is that it allows our design strategy to focus directly on respondent composition with respect to target population. In addition, this strategy provides statistically justified selection probability during the sampling stage.

One key element of improving the respondent representativeness is to balance multivariate distributions between respondents and the target population. A natural choice for balancing multivariate distribution is to employ propensity score methods (Rosenbaum and Rubin, 1983). When using propensity scores to design an observational investigation, the goal is to match 'treated subjects' with 'control subjects' having similar measured covariates, and in turn reduce bias from not comparable groups, and increase precision in estimating the 'treatment effect' (Rubin, 2002, 2007; Austin, 2009; Hahn, et. al., 2011). This approach, however, is not a good fit with survey design, where a priori matching of nonrespondents with comparable replacement subjects is not feasible. Response status is not known prior to the survey and advance information required to match is limited for both respondents and nonrespondents. Hence, the nonresponse problem in surveys undermines the utility of a propensity score matching approach that is a common approach in reducing bias in other observational studies.

In the survey research context, propensity score methods are commonly used after the survey is complete to estimate the propensity of response for the purposes of developing a nonresponse weighting adjustment. The success of these methods in reducing bias depends on the strength of association between auxiliary nonresponse information and the response propensity (Little and Vartivarian, 2003). A major difference between this research and the propensity weighting adjustment is that the conventional nonresponse adjustments are implemented at the conclusion of a survey and primarily focus on the use of nonrespondent information such as frame data, registry data, and proxy measures collected by interviewers (Wun and Ezzati-Rice, 2007; Groves, et. al., 2007; Kreuter et. al., 2010). The emphasis of the proposed method is on the proactive strategy at the survey design and data collection stage; on respondent representativeness rather than response rate; on prevention rather than adjusting for differential nonresponse; and on dynamically benchmarking based on external

sources instead of confined to the nonrespondent information.

One may question the advantages of sampling from new sample replicates instead of exclusively pursuing nonrespondents from original samples. The benefits are twofold. First, nonrespondents may reach phase-capacity and require a different protocol in order to prompt them to respond, such as higher incentives, different data collection modes, more experienced interviewers, and different refusal conversion techniques. These strategies introduce potential measurement error and nonsampling bias. In comparison, in the proposed responsive method all that needs to be changed is the sample allocation in the successive data collection phases to enhance the representativeness, which minimizes the operational complexity and avoids introducing the un-intended survey errors. The second benefit is that, there may be frame deficiencies, even if most respondents responded, the weighting procedures such as raking, poststratification, and calibration may have to do heavy lifting at the weighting stages which increases the weight variability.

By continuously benchmarking to the target population during the survey, the goal of raking and post-stratification is incorporated into the data collection process, which may minimize the scale of corrective weighting. Also, given the unknown status of nonresponse mechanism, more representative respondents require less nonresponse adjustment and hence less bias and variance. An added benefit of benchmarking to the target population using propensity scores is the preservation of the multivariate structure in terms of balance between benchmark and the new survey. Such a multivariate balance provides a more representative respondent pool with respect to both the marginal and joint distributions, which also improves the precision of regression coefficients estimated using the sample, such as in prediction and imputation models.

In section 2, we first describe the data used and define the notation that is used throughout the article. In section 3, we explain our methods, first we presents a simple univariate example to demonstrate the principle of our method. We then derive the sampling rate for an underrepresented population, modeling on propensity score along with strategies for model diagnostics. Section 4 provides the simulation results of the proposed design. In section 5 we conclude with some discussion of limitations, the advantages and disadvantages of the proposed strategy comparing to weighting-based bias adjustment, along with areas for future research.

2. Data used and definitions of conditions

We conduct a hypothetical responsive design using NHIS as benchmark and BRFSS as the new survey. NHIS is a high quality face-to-face survey that shares many common background and outcome variables with BRFSS. Therefore, the pair of NHIS and BRFSS is an ideal example of a benchmark and a new survey.

2.1 The NHIS and the BRFSS

The NHIS, initiated in 1957, is a cross-sectional continuing survey of the National Center for Health Statistics (NCHS). NHIS monitors trends in illness and disability of the civilian, non-institutionalized, household population of the United States. NHIS is the principle source of information about the health of the United States population. Aside from collecting health characteristics, NHIS includes many demographic and socioeconomic characteristics.

The BRFSS is an on-going telephone health survey where data are collected monthly by each of 50 States and district of Columbia. BRFSS tracks health conditions and risk behaviors in the US since 1984. Both the NHIS and the BRFSS are multi-purpose health

surveys that share many health related questions in common. Being a household face-to-face interview survey with a relatively high overall response rate (80.1% conditional response rate and 65.4% final response rate in 2009), NHIS is considered superior than the BRFSS (34.9% median response rate in 2009), a telephone survey, in general (Nelson, et al., 2002). However, the BRFSS being the world's largest health survey has sample sizes of more than six times larger that of the NHIS. The data provided by a survey with such a large sample size trumps other shortcomings.

Public-use micro-data from the NHIS and the BRFSS are used here. NHIS includes subjects from 50 states and DC whereas BRFSS include subjects from 50 states and US territory, such as Guam, Puerto Rico and Virgin Island. For comparability, subjects from US territory are excluded from BRFSS data. Also, subjects who are younger than 18 years old are excluded. Common variables (some recoded) from the two surveys are used in illustrating and evaluating the proposed responsive design method. These variables are those commonly used in the post-survey weighting adjustment for household survey samples and includes demographic and socio-economic characteristics such as age, sex, race/ethnicity, marital status, education, health insurance coverage, and household income, etc. These variables are chosen to demonstrate that a better represented respondent pool can be obtained by proposed sampling strategy by using the same information that are commonly applied at the corrective weighting steps in a traditional design.

2.2 Comparison of 2009 NHIS and 2009 BRFSS under conventional design

Prior to the comparison, we first evaluated the covariates shared by both 2009 NHIS and 2009 BRFSS. Covariates in common between NHIS and BRFSS are listed in Table 1, including basic demographic and background information. These covariates are also common to many other large government surveys. Also listed in Table 1 are summary statistics for these covariates, such as means for continuous variables and proportions of categorical variables. These summary statistics are computed using the 2009 NHIS and the 2009 BRFSS in the original observed data, which are available on the corresponding websites with public-use micro data (`http://www.cdc.gov`).

To illustrate the departure of respondent composition from target population that can occur in survey practice, the column titled 'NHIS' in table 1 listed weighted NHIS estimates which serve as a benchmark for the target population. The next four columns consist of unweighted BRFSS covariates which represents estimates of a 'current survey' that implemented conventional survey design.

Four BRFSS columns, each provides the descriptive statistics of observed data accumulated to the corresponding data collection phases. For example, the column titled 'BRFSS Phase 4' represents BRFSS data accumulated from the first to the fourth phase of data collection. There are obvious and marked imbalances for several covariates, including age, gender, education, and Hispanic origin, beginning at phase 1 and continuing across all phases. These differences suggest differential nonresponse among subgroups, which indicates a potential for nonresponse bias in survey estimates. These differences cannot be completely attributed to sampling error.

The comparison in Table 1 illustrates the departure of BRFSS from the target population in an univariate fashion, one variable at a time.

Table 1: Common variables and descriptive statistics from NHIS 2009 and BRFSS 2009

Variable	Values	NHIS[1]	BRFSS Phase 1	BRFSS Phase 2[2]	BRFSS Phase 3	BRFSS Phase 4	Absolute[3] differences
Region	Northeast	18.0	20.3	18.8	18.7	18.5	0.5
	Midwest	23.1	23.2	23.3	23.9	24.0	0.9
	South	36.2	32.4	31.6	31.7	31.4	4.8
	West	22.7	24.1	26.2	25.7	26.1	3.4
Age	(in years)	46.1	55.4	55.5	55.7	55.9	9.8
Gender	Male	48.3	38.5	38.2	38.0	38.1	10.2
	Female	51.7	61.5	61.8	62.0	61.9	10.2
Marital	Married	55.2	56.8	56.4	56.4	56.4	1.2
Status	Widowed	5.8	14.0	14.0	14.1	14.1	8.3
	Divorced	8.5	13.5	13.6	13.7	13.7	5.2
	Separated	2.1	2.0	2.0	2.0	2.0	0.1
	Never	21.1	13.4	13.6	13.6	13.5	7.6
	Unknown	7.4	0.3	0.3	0.3	0.3	7.1
Education	$\leq$ Kindergarten	0.4	0.1	0.1	0.1	0.1	0.3
	Grade 1-8	4.8	2.9	3.1	3.1	3.1	1.7
	Grade 9-11	9.5	6.2	6.2	6.1	6.1	3.4
	G12/GED	28.5	30.4	30.2	30.1	30.0	1.5
	College 1-3 yrs	30.0	26.8	27.0	26.9	26.9	3.1
	College $\geq$4 yrs	26.8	33.5	33.4	33.6	33.8	7
Hispanic	Yes	13.8	5.7	6.1	6.0	6.1	7.7
	No	86.2	94.3	93.9	94.0	93.9	7.7
Race	White only	80.8	84.8	84.7	84.5	84.4	3.6
	Black only	12.1	7.9	7.9	8.0	8.1	4
	AIAN[4]	0.8	1.7	1.7	1.6	1.6	0.8
	Aisan only	4.7	1.6	1.7	1.7	1.7	3
	Other single race	0.3	2.3	2.3	2.3	2.3	2
	Multiracial	1.3	1.8	1.8	1.8	1.8	0.5
Health	Yes	84.6	89.7	89.5	89.5	89.6	5
Insurance	No	15.4	10.3	10.5	10.5	10.4	5
Income	$<$10 K	5.6	5.1	5.1	5.2	5.2	0.4
(Annual	10K-15K	4.5	5.8	5.9	6.0	6.0	1.5
Household	15K-20K	4.9	7.8	7.9	7.9	7.9	3
income from	20K-25K	6.2	9.5	9.7	9.7	9.7	3.5
all sources)	25K-35K	10.4	12.1	12.1	12.1	12.2	1.8
	35K-50K	14.3	15.7	15.6	15.6	15.5	1.2
	50K-75K	18.6	17.0	16.6	16.5	16.5	2.1
	$\geq$ 75K	35.6	27.1	27.1	27.1	27.1	8.5

[1]NHIS: weighted, serves as benchmark [2]BRFSS: unweighted, serves as current survey. Phase 2 represents cumulative data, including data from both phase 1 and 2. Similarly, phase 3 and phase 4 columns reflect cumulative data up to the corresponding phases. [3]Absolute differences comparing phase 4 (overall sample) to benchmark. [4]AIAN : American Indian or Alaska Natives.

3. Methods

3.1 The design

The proposed design starts with a current survey that is planned under a chosen survey design. The current survey is conducted based on the initial design until the collection of sufficient number of respondents or the completion of the first phase data collection. For example, a one-year survey may be divided into 4 phases of data collection where each phase lasting 3 months represents a replicate sample of the target population.

Concurrent to the planning of a current survey is the identification of a benchmark, which might be a census or a high quality survey that captures a target population. The benchmark and the current survey share common covariates (denoted as $\boldsymbol{Z}$) which are available as frame information in the establishment surveys and household surveys. For example, in an establishment survey these covariates might be hospital beds, hospital size, etc. In an address-based household survey these covariates might be census hard to count score, census block level median household income, gender ratio, etc.

With a benchmark in mind, once sufficient data is collected, the collected data is compared to the benchmark and a responsive strategy is developed to tailor the design. At the completion of the first phase of data collection, data from the current survey is compared to the data from the benchmark with respect to the common covariates to evaluate the balance. The balance is defined as the similarity of benchmark $\boldsymbol{Z}$ distribution ($P_B(\boldsymbol{Z})$) and current survey $\boldsymbol{Z}$ distribution ($P(\boldsymbol{Z})$). A balance between the two surveys is achieved when $P(\boldsymbol{Z}) = P_B(\boldsymbol{Z})$.

While imbalance exists, subjects that are underrepresented with respect to $\boldsymbol{Z}$ will be sampled in a higher rate at the next phase of data collection, whereas subjects that are overrepresented will be sampled in a lower rate. This required adjustments to sampling rates for subpopulations and can be derived under the principle described in the univariate example described in the next section.

3.2 Numerical example on a univariate case

Before the formal derivation of the method, we use a simple univariate example based on gender distribution to illustrate the core concept. Suppose the gender distribution in the target population is 50% males and 50% females. Suppose in a survey with two phases of data collection, the desired total sample size is 100 respondents. After the first phase of data collection, 60 respondents (60%) are obtained where 63% are female and 37% are male. In comparison to the target population of 50% female, the obtained female distribution, 63%, is over-represented. In the second phase of data collection we would like to undersample females and oversample males. This paper propose a sampling method to obtain the remaining 40 respondents from the second sample replicates, such that the final gender distribution resembles that of the target population as closely as possible.

To write out the above problem algebraically, consider the female distribution under current objective. That is, the target female percentage of 50% is achieved by combining the female percentage from the first phase data collection, 0.6×0.63, and the second phase data collection, $0.4 \times F$, where F represents the desired female distribution for the second phase of data collection. Assuming the response rates for women will not differ between two sample replicates, we can write

$$0.4 \times F + 0.6 \times 0.63 = 0.5 \tag{1}$$

In this univariate case, we can solve for F numerically, $F = (0.5 - 0.6 \times 0.63)/0.4$. Now we use mathematical notation to rewrite formula (1).

Let Z denote gender, and $P(Z)$ denote the gender distribution. Let $P_B(Z)$ denote the benchmark gender distribution. Furthermore, let π denote the proportion for the sample remaining to be interviewed, i.e. phase 2 sample, and hence $1-\pi$ is the sample proportion of obtained subjects at the first phase of data collection. F, the desired female distribution in (1), is now the desired gender distribution conditioning on subjects being selected in the sample, denoted as $P(Z|sampled)$. Using these notations, equation (1) can be rewritten as

$$\pi \times P(Z|sampled) + (1-\pi) \times P(Z) = P_B(Z) \tag{2}$$

Since $P(Z)$, $P_B(Z)$, and π are known, we can solve for $P(Z|sampled)$. By rearranging the terms in (2), we have $P(Z|sampled) = (P_B(Z) - (1-\pi)P(Z))/\pi$. However, what we really want to know is how to select samples that would most likely to result in the desired $P(Z|sampled)$. That is, we actually want to know the sampling rate conditioning on the gender distribution, which is $P(sampled|Z)$.

The association between $P(sampled|Z)$ and $P(Z|sampled)$ can be expressed as

$$P(sampled|Z) = \frac{P(Z|sampled)P(sampled)}{P(Z)} \tag{3}$$

Putting formulas (2) and (3) together, we have

$$P(sampled|Z) = \{\frac{P_B(Z)}{P(Z)} - (1-\pi)\}\frac{P(sampled)}{\pi} \tag{4}$$

In our numerical example, $\pi = 0.4$, $P(Z) = 0.63$, and $P_B(Z) = 0.5$. Plugging into (4), we obtain $P(sampled|Z) = [(0.5/0.63) - 0.6] \times P(sampled) \times 0.4$. Now, $P(sampled)$ is not directly known but can be computed. Since $P(sampled|Z)$ is a probability, bounded by 0 and 1. We can write

$$\max_Z \{\frac{P_B(Z)}{P(Z)} - (1-\pi)\}\frac{P(sampled)}{\pi} \leq 1$$

To find a solution for $P(sampled)$, we set

$$P(sampled) = \frac{\pi}{\max_Z \{\frac{P_B(Z)}{P(Z)} - (1-\pi)\}}$$

3.3 Computing sampling rate

When Z is univariate, such as the gender example, computing $P_B(Z)/P(Z)$ is straightforward. However, in most situations Z is a vector consisting of many covariates, including both main effects (i.e. marginal distributions) and interactions (i.e. conditional distributions). For example, for a respondent pool that resembles the target population with covariates listed in table 1, $Pr(\boldsymbol{Z})$ would include both the marginal and the conditional distributions of those covariates. Specification of the joint distribution of a vector consisting of continuous and categorical variables is complex and difficult. The derivation would be considerable simplified if Z could be univariate. Fortunately, reducing multivariate distribution of $\boldsymbol{Z}$ to a scalar can be achieved by the propensity score methods (Rosenbaum and Rubin, 1983).

Therefore, if we consider $P(Z)$ to be the propensity score density of the current survey up until phase k, and $P_B(Z)$ to be the propensity score density of the benchmark survey, then $P_B(Z)/P(Z)$ becomes the ratio of propensity score density and the computation of this density ratio is greatly simplified.

The propensity score model is estimated using logistic regression. Covariates common to both the benchmark and the current survey are included in the regression model. The outcome variable of survey membership is coded 1 for a subject "being in the benchmark survey". The estimated propensity score is the predicted probability of "being in the benchmark survey" derived from the fitted logistic regression model. The data structure for propensity score modeling can be visualized in Figure 2. Specifically, let M_i denote the survey membership indicator, where $M_i = 1$ if subject i is from benchmark survey and $M_i = 0$, otherwise. Furthermore, let Z_i denotes the background variables for subject i. After the first phase of data collection, the propensity score is estimated using M and data Z_B and $Z_R^{(1)}$. More formally, the propensity score, $e_i^{(1)} = P(M_i = 1|Z_i)$, is the probability of "being in the benchmark survey" for subjects with background characteristics of Z_i.

4. Simulation Results

4.1 Simulation Study 1: A multivariate Normal Situation

A first simulation study demonstrates how the ratio of propensity score densities for a current survey and benchmark can be used to adjust the relative sampling rate for population members in subsequent phases of the responsive design. We simulated a benchmark survey consisting of N subjects and p covariates where p covariates, denoted as $\boldsymbol{Z}_B$ with dimension of $N \times p$, have multivariate normal distribution. Similarly, a current survey is simulated consisting of $n^{(k)}$ respondents at phase k with the same p covariates, denoted as $\boldsymbol{Z}$, with dimension of $n^{(k)} \times p$, and $\boldsymbol{Z}$ also has multivariate normal distribution.

Let M denote the study label where the benchmark survey sample members have $M = 1$ and the current survey sample respondents have $M = 0$. We fit a propensity score model using M and Z. The estimated propensity score $e_i(Z)$ for subject i is the probability of subject i with characteristics Z in the current survey. The ratio of propensity score densities between the benchmark survey, $P_B(e)$, and the current survey, $P^{(k)}(e)$, is denoted $w^{(k)}$. $w^{(k)}$ is the relative sampling rate for the $k+1$ data collection phase. We iterate the process 4 times to simulate the 4 phases of data collection. The first row of density function graphs in Figure 4 bottom row compares the empirical propensity score distribution for the benchmark survey and the current survey at each successive data collection phase.

It is evident that the propensity score distribution of the two surveys converges further after each data collection phase, indicating that the responsive design sampling rate adjustments successfully improves the respondent representativeness in the current survey after each data collection phase.

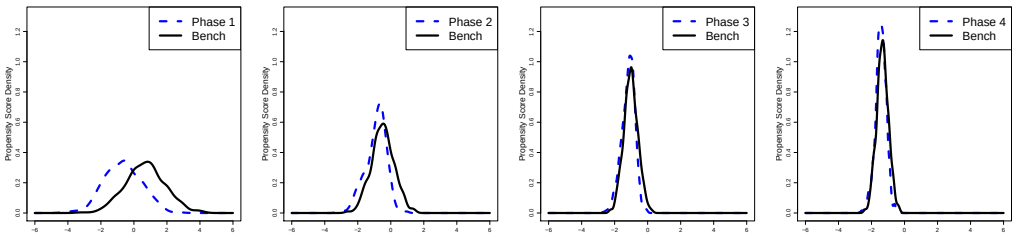

Figure 3: Estimated propensity distribution of multivariate normal example from the responsive design.

4.2 Simulation Study 2: Benchmarking BRFSS using NHIS

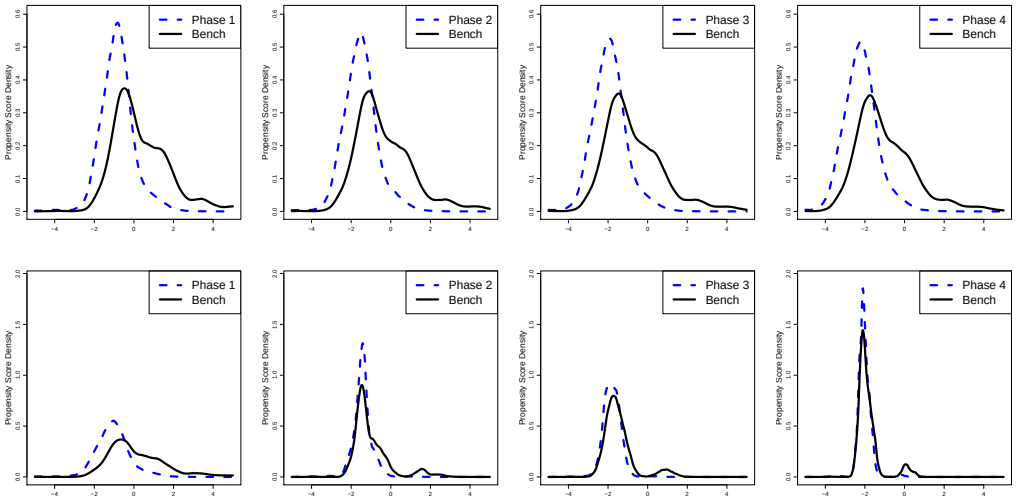

Figure 4: (Top row) Estimated propensity distribution of observed BRFSS under conventional design. (Bottom row) Estimated propensity distribution on weighted NHIS (Benchmark) and unweighted BRFSS from the hypothetical responsive design.

Before the implementation of the proposed responsive design, we first illustrate the departure of BRFSS from the target population in a multivariate fashion under the traditional design. We compare the respondent samples from both NHIS and BRFSS using their propensity score density function. Variables used in the propensity score model included geographical region, age, sex, Hispanic origin, race, marital status, education, income and health care coverage.

As shown in the time-specific plots provided in Figure 4 top row, the distributional differences between the benchmark survey (NHIS) and the current time BRFSS respondent sample slightly widened as the data collection progressed from Phase 1 to Phase 4. Without mid-survey intervention, the resulting characteristics of the respondent pool deviated from the target population further at each data collection phase.

Next, we implemented the responsive design strategy on a hypothetical survey based on data from the 2009 NHIS and the 2009 BRFSS. The weighted 2009 NHIS serves as the benchmark and the unweighted 2009 BRFSS serves as the current survey. This simulation study starts by concatenating the observed data from the 2009 BRFSS first quarter (January, February and March) to the 2009 NHIS data. The combined data is used to fit a propensity score model to estimate the probability of "being in the benchmark survey" using the covariates that are listed in table 2. The sampling rate is computed, and used to sample the phase 2 data. We sampled with replacement for the second, third and forth phases of respondents from the 2009 BRFSS data.

Table 2 illustrates the results of this hypothetical responsive design. Similar to table 1, the column of NHIS in table 2 listed summary statistics from the weighted 2009 NHIS, as a benchmark. Since the responsive sampling starts with the phase 2 data collection, the first phase data is the observed data from the 2009 BRFSS, which is same as the traditional design listed in table 1 under column "BRFSS phase 1", and thus it is omitted from table 2. The resulting data from the responsive design are listed under three BRFSS columns. As each phase of data accumulates, we see an overall incremental improvement on the distributions of covariates with respect to the benchmark.

The last two columns of table 2 listed absolute differences. One column listed the absolute differences between the benchmark and the responsive design, and the other column listed the absolute differences between the benchmark and the survey under traditional design, the observed data from the 2009 BRFSS, which is the same as the last column on table 1. In all variables, the total absolute difference is larger for the traditional design than for the responsive design. This result indicates that the covariate distributions from responsive design resembles those of the benchmark more closely than those of the traditional design.

The absolute differences from each covariate measure the balance of the two surveys in a univariate fashion. To assess this balance in a multivariate fashion, the propensity score density distribution was computed and the results are illustrated in the bottom row of figure 4. The findings of multivariate balance assessment echo that of the univariate absolute differences. The propensity score density distribution from the responsive design converges with the benchmark distribution at each sequence of data collection.

5. Discussion

Concerns about respondent representativeness have come to light in part due to declining response rates in recent years. The seriousness of this decline was revealed by the Pew Research Center in 2012 that reported an average 9% response rate for a typical telephone survey, which means an average of 91% nonresponse rate for a typical RDD survey. Household surveys funded by the U.S. government have also seen steadily increasing nonresponse rates. For example, from 1990 to 2004 initial contact nonresponse rates have approximately doubled for selected household surveys, including the Consumer Expenditure Quarterly (from 12% to 23.3%), the Current Population Survey (from 5.7% to 10.1%), and the Survey of Income and Program Participation (from 7.3% to 14.9%).

Many strategies directed at increasing the response rate have implications leading to increasing costs, such as those associated with additional interviewer training and various incentive programs. Moreover, the association between overall response rates and nonresponse bias has been shown to be inconsistent (Keeter et al., 2000; Curtin et al., 2000; Merkle and Edelman, 2002; Groves, 2006; Groves and Peytcheva, 2008).

This study examined the effect of the implementation of a responsive design on enhancing the representativeness of the observed respondent sample. The 2009 NHIS and the 2009 BRFSS publicly available micro-data are used for illustration. We demonstrate the feasibility of a responsive design that uses propensity score derivatives and a high quality benchmark in aiding sampling decisions in a large-scale multi-phase survey (e.g. BRFSS). Further, we show the capacity to obtain a more representative respondent pool while controlling the cost and following the same timeline of a corresponding conventional design. Briefly, targeting underrepresented populations up to the current phase using a sampling rates adjusted by propensity scores has been shown to be suitable to implement in the proposed design.

The proposed responsive design improves respondent representativeness by sequentially evaluating the resemblance between the current respondent pool and external benchmark in terms of distributions of subject characteristics, instead of targeting the response rate and number of completes. The strategy is simple and intuitive, easy to implement, multivariate in nature (can incorporate both main effects and interactions), and then method applies to the entire sample, rather than individual cases. Our strategy is superior in two ways: 1) survey respondents become more representative of the target population in a multivariate fashion which preserves the variance-covariance structure of the sample composition, and 2) more representative respondent pool minimizes the need for final corrective weighting adjustments, leading to less nonresponse bias and less weight variability.

Table 2: Comparison of summary statistics between NHIS and BRFSS 2009 in the simulated data under responsive design

Variable	Values	NHIS[1]	Responsive BRFSS Phase 2	Responsive BRFSS Phase 3[2]	Responsive BRFSS Phase 4	Responsive Design Absolute Difference[3]	Traditional Design Absolute Difference[3]
Region	Northeast	18.0	13.2	17.7	17.8	0.2	0.5
	Midwest	23.1	22.0	20.5	23.3	0.2	0.9
	South	36.2	40.5	30.4	36.6	0.4	4.8
	West	22.7	24.3	31.4	22.3	0.4	3.4
Age	(in years)	46.1	44.4	50.5	55.7	9.6	9.8
Gender	Male	48.3	49.3	42.7	47.0	1.3	10.2
	Female	51.7	50.7	57.3	53.0	1.3	10.2
Marital Status	Married	55.2	61.5	59.1	58.9	3.7	1.2
	Widowed	5.8	4.0	8.0	8.6	2.8	8.3
	Divorced	8.5	7.4	11.1	10.5	2.0	5.2
	Separated	2.1	2.9	2.2	2.2	0.1	0.1
	Never	21.1	23.5	18.4	18.7	2.4	7.6
	Unknown	7.4	0.7	1.3	1.1	6.3	7.1
Education	$\leq$ Kindergarten	0.4	0.3	0.5	0.4	0.0	0.3
	Grade 1-8	4.8	5.5	6.5	4.9	0.1	1.7
	Grade 9-11	9.5	10.6	9.6	8.6	0.9	3.4
	G12/GED	28.5	31.3	24.6	30.1	1.6	1.5
	College 1-3 yrs	30.0	31.1	24.0	28.9	1.1	3.1
	College $\geq$4 yrs	21.3	33.5	34.9	26.6	5.3	7
Hispanic	Yes	13.8	16	21.4	12.4	1.4	7.7
	No	86.2	84	78.6	87.6	1.4	7.7
Race	White only	80.8	69.5	81.3	79.8	1.0	3.6
	Black only	12.1	21.8	8.4	12.8	0.7	4
	AIAN[4]	0.8	0.7	7.8	1.2	0.4	0.8
	Aisan only	4.7	6.8	0.07	3.7	1.0	3
	Other single race	0.3	0.02	1.6	1.1	0.8	2
	Multiracial	1.3	1.8	1.2	1.4	0.1	0.5
Health Insurance	Yes	84.6	86.3	85.4	86.3	1.7	5
	No	15.4	13.7	14.6	13.7	1.7	5
Income (Annual household income from all sources)	$<$10 K	5.6	6.9	5.5	5.1	0.5	0.4
	10K-15K	4.5	4.4	4.7	5.0	0.5	1.5
	15K-20K	4.9	4.9	5.4	6.1	1.2	3
	20K-25K	6.2	5.5	6.8	7.6	1.4	3.5
	25K-35K	10.4	9.1	11.8	10.3	0.1	1.8
	35K-50K	14.3	12.7	17.5	15.0	0.7	1.2
	50K-75K	18.6	18.3	17.8	19.7	1.1	2.1
	$\geq$ 75K	35.6	38.1	30.4	31.3	4.3	8.5

[1]NHIS: weighted, serves as benchmark [2]BRFSS: unweighted, serves as current survey. Phase 2 represents cumulative data, including data from both phase 1 and 2. Similarly, phase 3 and phase 4 columns reflect cumulative data up to the corresponding phases. [3]Absolute differences comparing phase 4 (overall sample) to benchmark. [4]AIAN : American Indian or Alaska Natives.

Responsive designs in the current literature describe construction of metrics that are informative to cost and survey error and use these for continuous monitoring. Metrics such as response rate across demographic groups, sampling strata and geographic areas are evaluated and inform plans aimed at increasing participation to achieve a desired sample size for analysis. Models are then constructed using auxiliary variables such as census geocoding to estimate the response propensities, optimal times to call, and identifying better-performing interviewers.

These examples are typical in current responsive design where balancing is between the respondent pool and the original sample, and where complexity in balancing increases rapidly when the number of covariates increases. Our method focuses on balancing to the target population and the balancing approach can easily incorporates both main effects and interactions. Our univariate example demonstrated the simplicity of deriving the proposed sampling rate. As the number of covariates increase, the sampling rate computation becomes manually un-attainable. And the methods section illustrates the proposed strategy to reduce the multivariate nature of multiple covariates to a single propensity score and to derive the sampling rate and demonstrated its theoretical justification.

One may question the need to improve respondent representativeness since nonresponse weighting adjustments correct for nonresponse bias. However, standard nonresponse adjustment correct for nonresponse bias only when the missing data mechanism is ignorable, but the literature does not always support this assumption. Aside from unverifiable assumption, the standard nonresponse adjustment procedures relies heavily on nonrespondent information that is often limited to derive the weighting adjustments.

The strategy of a responsive design illustrated in this paper offers an alternative solution to reduce nonresponse bias by first improving the respondent representativeness and consequently minimizes corrective weighting adjustments, increasing the precision of the survey estimates.

The proposed design may increase survey cost in certain situations. For example, if the underrepresented population is hard-to-reach and requires more resources to obtain a response. This is not necessarily a limitation of the design since such incurred extra cost would be the same for traditional survey design. Another limitation is the availability of a benchmark. A benchmark could be information from the Census or a similar survey conducted previously. For example, for recurrent surveys, one may use the weighted estimates from previous years as benchmark for the current survey.

Furthermore, many household surveys lack a comprehensive frame information which limits the use of the proposed sampling strategy. This strategy can be further improved by predicting the survey variables by using models that fitted to the frame information. Specifically, recall figure 2. One can use models fitted to Z_B and $Z^{(1)}$ and models fitted to X_B and $X^{(1)}$ to predict $X^{(2)}$ before data collection. Once phase II data is collected, $X_R^{(2)}$ becomes known and can be used to validate the prediction model and modify it accordingly. The implication of this step is that, if the interest is to sample subjects with unobserved characteristics, such as rare disease or condition, this sampling strategy could potentially obtain a higher number of the desired sample with a controlled budget, as compared to a large scale screening phase.

REFERENCES

Austin PC. (2009), "Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched sample." *Statistics in Medicine*, 28: 3083–3107.

Barclay S, Todda C, Finlayb I, Grande G, Wyattc P. (2002), "Not another questionnaire! Maximizing the response rate, predicting non-response and assessing non-response bias in postal questionnaire studies of GPs." *Family Practice*, 19 (1): 105–111.

Biemer, P.P., Peytchev, A. (2012), "Census Geocoding for Nonresponse Bias Evaluation in Telephone Surveys An Assessment of the Error Properties." *Public Opinion Quarterly*, 76 (3):432–452.

Bootsma-van der Wiel A, van Exel E, de Craen A.J.M, Gussekloo J, Lagaay A.M, Knook D.L., Westendorp R.G.J. (2002), "A high response is not essential to prevent selection bias: Results from the Leiden 85-plus study." *Journal of Clinical Epidemiology*, 55(11): 1119–1125.

BRFSS 2009 public use microdata files available for download at `http://www.cdc.gov/brfss/technical_infodata/surveydata/2009.htm` (accessed on February, 2012)

Curtin R, Presser S, Singer E (2000), "The Effects of Response Rate Changes on the Index of Consumer Sentiment." *The Public Opinion Quaterly*, 64(4): 413–428.

D'Agostino Jr RB. (1998), "Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group." *Statistics in Medicine*, 17: 2265–2281.

de Leew ED, de Heer W. (2002), "Trends in household survey nonresponse: a longitudinal and international comparison", *Survey Nonresponse*, ed. Robert M. Groves, Don A. Dillman, John L. Eltinge, and Roderick J. A. Little, pp. 41–54. New York: Wiley.

Groves R, Couper M. (1998), *Nonresponse in Household Interview Surveys.* New York: John Wiley.

Groves R, Singer E, Corning A. (2000), "Leverage-Saliency theorey of survey nonresponse: Description and an illustration." *Public Opinion Quarterly*, 64(3): 299–308.

Groves RM, Heeringa, SG. (2006), "Responsive design for household surveys: tools for actively controlling survey errors and costs." *J. R. Statist. Soc. A*, 169(3): 439–457.

Groves RM (2006), "Nonresponse rate and nonresponse bias in household surveys." *The public Opinion Quarterly*, 70(5): 646–675

Groves R., Wagner J, Peytchev E. (2007) "Use of interviewer judgments about attributes of selected respondents in post-survey adjustment for unit nonresponse: An illustration with the National Survey of Family Growth." In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*.

Groves R, Peytchev E. (2008) "The impact of nonresponse rates on nonresponse bias." *The Public Opinion Quarterly*, 72(2): 167–189.

Hahn J, Hirano K, Karlan D. (2011), "Adaptive experimental design using the propensity score." *Journal of Business and Economic Statistics*, 29(1): 96–108.

Keeter S, Miller C, Kohut A, Groves R, Presser S. (2000), "Consequences of Reducing Nonresponse in a National Telephone Survey." *The Public Opinion Quarterly*, 64(2), 125–148

Kreuter F, Couper M, Lybery L. (2010) "The use of paradata to monitor and manage survey data collection." In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 282–296.

Lee S, Belin T, Brick M, Wang Z (2009), "Responsive design for random digit dial surveys using paradata." *ESRC National Centre for Research Methods*, The application of Paradata in Social Surveys One-Day seminar, available at `www.natcen.ac.uk/ncrm-paradata-network/docs/NCRM_Lee.pdf`

Little R, Vartivarian S (2003), "On weighting the rates in non-response weights." *Statistics in Medicine*, 22(9): 1589–1599.

Little R, Vartivarian S (2005), "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology*, 31(2): 161–168.

Merkle D, Edelman M (2002), "Nonresponse in exit polls: A comprehensive analysis." In *Survey nonresponse*, ed. Robert M. Groves, Don A. Dillman, John L. Eltinge, and Roderick J. A. Little, pp. 243-258. New York: Wiley.

Nelson D, Powell-Griner E, Town M, Kovar M (2002), "A Comparison of National Estimates From the National Health Interview Survey and the Behavioral Risk Factor Surveillance System." *American Journal of Public Health*, 93(8): 1335–1341.

NHIS 2009 public use microdata files available for download at `http://www.cdc.gov/nchs/nhis/nhis_2009_data_release.htm` (accessed on February, 2012)

Pew research report (2012), "Assessing the Representativeness of Public Opinion Surveys", `http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/`, accessed on July, 2012.

Peytchev A, Rosen J, Riley S, Murphy J, Lindblad M. (2010a), "Reduction of Nonresponse Bias in Surveys through Case Prioritization." *Survey Research Methods*, 4(1): 21–29.

Peytchev A. (2010b), "Responsive design in telephone survey data collection." In *Proceedings of household survey nonresponse workshop 2010*, available at `http://www.nonresponse.org/index.php?fl=2&lact=1&bid=621&avtor=509&parent=3` (accessed on July, 2012)

Rosenbaum PR, Rubin DB. (1983), "The central role of the propensity score in observational studies for causal effects." *Biometrika*, 70, 41–55.

Royall R, J Herson. (1973), "Robust Estimation in Finite Populations I." *Journal of the American Statistical Association*, 68(344): 880–889.

Rubin DB. (2002), "Using propensity scores to help design observational studies: Application to the tobacco litigation." *Health Serv. and Outcomes Res. Methodol.*, 2: 169–188.

Rubin DB. (2007), "The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials." *Statistics in Medicine*, 26(1): 20–36.

Schouten B, Cobben F. (2007), " R-indexes for the comparison of different fieldwork strategies and data collection modes." Discussion paper 07002, Voorburg/Heerlen, The Netherlands: Statistics Netherlands.

Schouten B, Shlomo N, Skinner C. (2010), "Indicators for monitoring and improving representativeness of response." *RISQ project*, accessible at `http://eprints.soton.ac.uk/158353/`

Wun LM, Ezzati-Rice TM, Diaz-Tena N, Greenblatt J. (2007), "On modelling response propensity for dwelling unit (DU) level non-response adjustment in the Medical Expenditure Panel Survey (MEPS)." *Statistics in Medicine*, 26(8): 1875–1884.