The Use of Indicators to Assess the Quality of Business Survey Returns During Data Collection

Daniel Whitehead, Broderick Oliver, Yarissa González[1]
U.S. Census Bureau; 4600 Silver Hill Road; Washington, DC 20233
Daniel.Whitehead@census.gov

Abstract

Data collection efforts often focus on maximizing survey response. However, increasing the response rate does not necessarily improve the quality of the estimates. For example, the respondents and nonrespondents might differ systematically on key survey characteristic(s). In this case, without successful data collection strategies to obtain data from underrepresented subpopulations, additional collection efforts may not improve the quality of the estimates. Using empirical data from two surveys, we examine two indicators, each measuring a separate property of the respondent sample: the R-indicator, which measures deviation from missing-completely-at-random; and the balance indicator, which measures the deviation of the respondent-based mean from the full sample mean for selected items. Examined in conjunction with the weighted volume response rate, which estimates the population coverage, these two indicators signal when the response set has stabilized but is not a random subset of the full sample. Thus, the current data collection strategy is at phase capacity, and new collection strategies -- targeted to specific subpopulations, are needed to achieve a balanced response.

Key Words: adaptive design, data collection, nonresponse bias, representativeness, response indicators, response rate

1. Introduction

Sample surveys are conducted to estimate characteristics of a specific population at a particular point in time. A good (i.e., "representative") sample will reproduce the characteristics of interest in the population, as closely as possible. To achieve this goal, the sampled units are selected via a probability-sampling plan where every unit of the population has a known probability of being included in the sample (Lohr, 1999).

The ideal response rate is 100 percent since surveys are designed with measurable sampling errors and unbiased estimates for a given sample size (Groves, 2006). However, this is just an ideal as most surveys suffer from some level of nonresponse. The main problem with nonresponse is the potential for biased population estimates. If there is a relationship between which units respond and the survey variables, then a systematic bias may be introduced into the estimates, lowering their accuracy (Lohr, 1999; Peytcheva and Groves, 2009).

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

A high response rate does not necessarily mitigate nonresponse bias. The composition of the response set is also important (Peytcheva and Groves, 2009). For this reason, a growing number of researchers have proposed alternative response indicators that incorporate both unit response and a measure of similarity between the response set and the full sample (see Wagner, 2012). Two such indicators that we examine in this paper are the *R-indicator* proposed by Schouten et al. (2009) and the *balance indicator* proposed by Särndal (2011). These indicators measure the degree to which the response set is similar to the full sample with respect to auxiliary variables or paradata available to all units on the frame.

The purpose of this research is to examine the potential use of these indicators during the data collection period for business surveys conducted by the U.S. Census Bureau to determine when the collection has reached phase capacity, meaning that the units that continue to respond are not improving the composition of the response set. When phase capacity is reached, one would expect that these indicators would either stabilize (if the same cross-section of unit type continues to respond) or decrease (if the disparity in response between domains continues to increase). At this point, as part of an adaptive design, new collection strategies that target specific subdomains are needed to increase the response of the underrepresented units (Groves and Heeringa, 2006; Wagner, 2012; Kreuter, 2013 (Ed.)). Schouten et al. (2011) introduce the use of partial R-indicators for achieving this purpose while Wagner (2012) discusses the idea of using subdomain response rates as a nonresponse analysis tool. For this paper, we chose the latter, which would be easier for the subject-matter analysts to interpret.

While the R-indicator and balance indicator both provide a measure of representativeness for the response set, the level of response must also be considered, particularly for business surveys where the populations are skewed and some units contribute significantly more to the estimates of interest than others. For this reason, we also examine the *weighted volume response rate* (WVRR), a weighted measure that approximates the proportion of the frame population that responds to the survey. We retrospectively apply the R-indicator and balance indicator in conjunction with the WVRR at regular intervals during the course of the data collection period of two of our business surveys to assess when, or if, the response set has reached phase capacity. We then examine response rates at the subdomain level to determine why or why not.

2. Indicators of Survey Response for U.S. Census Bureau Business Surveys

Unlike their household counterparts, business populations are often highly skewed. For example, the estimate of an industry total for a characteristic of interest may derive the majority of its value from relatively few units. Therefore, business surveys at the U.S. Census Bureau generally employ single stage samples with highly stratified designs that include these large units with certainty. The remaining units are sampled. Nonresponse from any of these large (certainty) units can result in biased estimates (Oliver and Thompson, 2013). Because of the importance of these certainty units to the estimates, analysts make a concerted effort to obtain their data during data collection.

When working with business surveys, “there is a need to distinguish between the survey (sampling) unit, the reporting unit, and the tabulation unit” (Oliver and Thompson, 2013, p.2). “The *survey unit* represents the entity selected via a probability sample from the frame. *Reporting units* are established by the sampled business to collect survey data. *Tabulation units* house the data for estimation.” (Thompson et al., forthcoming in 2014)

Business surveys at the U.S. Census Bureau compute both respondent level and item level response rates (Thompson et al., forthcoming in 2014). The *Unit Response Rate (URR)* is a respondent level response rate defined as the unweighted proportion of reporting units eligible for data collection or of unknown eligibility that respond to the survey (Thompson et al., forthcoming in 2014; Oliver and Thompson, 2013, p.3). The *Total Quantity Response Rate (TQRR)* is an item level response rate calculated for key data items based on information from tabulation units (Thompson et al., forthcoming in 2014). Both official measures are calculated within the U.S. Census Bureau's *Standard Economic Processing System* (StEPS) after data collection and editing are completed (Thompson and Oliver, 2012). The individual surveys establish required data items and minimally sufficient conditions to classify a unit as a respondent (Oliver and Thompson, 2013, p.2). However, in practice neither official measure can be tracked as collection occurs because response is not determined until after the data have been processed.

Because we wish to examine response at points in time throughout the collection period, we develop an **unofficial** unit response rate, which we refer to as a *Proxy Unit Response Rate (Proxy URR)*. This measure would only be possible to calculate during collection using data that is not yet fully edited, but because we are conducting a retrospective study we can track it in "real-time" using final data. We define the *Proxy URR* as the proportion of eligible reporting units (E) and units whose eligibility could not be determined (U) for which a questionnaire was received that was classified as a response (F) at a given moment of time within the statistical period:

$$Proxy\ URR = \frac{F}{E + U}$$

2.1 R-indicator

The R-indicator, developed by Schouten et al. (2009), provides a measure of the extent to which the units that respond to a survey are representative of the full sample in terms of response propensities derived from a set of auxiliary variable(s) or paradata available for all units sampled. To ensure that both the bias and variance of the characteristic means due to nonresponse are minimized, these auxiliary variables should be related to both survey response and the survey characteristics of interest (Little and Vartivarian, 2005). When this is true, obtaining response sets with high R-indicator values is desirable. Schouten (personal communication, June 13, 2014) argues that even if the auxiliary variables are not good predictors of response, the R-indicator can still be useful so long as the auxiliary vector is related to the variables of interest.

Schouten et al. (2009, p.103) characterizes a response set as being weakly representative with respect to the set of auxiliary variable(s) when the average response propensity for each subpopulation formed by the auxiliary variables is constant – hence equal to the overall response rate. In this case, the assumed response mechanism is missing-completely-at-random (MCAR) (Schouten et al., 2009, p.103). The more the response set deviates from MCAR, the less representative the data are of the full sample. The R-indicator ranges in value from 0 to 1 (see Notation and Formulae below). The higher the value, the more representative the response set.

Notation and Formulae:

i = 1, 2, 3…N represent a population of size N from which n sampled units were obtained via a given sampling plan

$w_i =$ sampling design weight for the i^{th} unit

$\hat{\rho}_i =$ an estimate of the true but unknown probability (ρ_i) that the i^{th} sampled unit will respond to the survey

$\hat{\bar{\rho}} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} \hat{\rho}_i w_i =$ the weighted sample average of the estimated response propensities and an estimator of $\bar{\rho}$, the mean of the true response propensities

$\hat{R}(\rho) = 1 - 2\sqrt{\frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i(\hat{\rho}_i - \hat{\bar{\rho}})^2}$ = estimate of the population R-indicator, $R(\rho)$.

2.2 Balance indicator

Similar to the R-indicator, Särndal's (2011) balance indicator measures the degree of similarity between the response set and the full sample in terms of auxiliary variables (or paradata) available to all sampled units. However, whereas the R-indicator measures "representativeness" based on response propensities, the balance indicator compares differences in weighted subdomain response rates across the auxiliary variables.

Särndal (2011, p.11) describes three separate versions of the balance indicator, all functions of the quadratic form, $\boldsymbol{D}' \sum_{\boldsymbol{S}}^{-1} \boldsymbol{D}$ (see Notation and Formulae). According to Särndal, the first version is similar "to 1 - R^2 in ordinary regression analysis" (2011, p.11) while the second version is always greater than or equal to the first and third version. Särndal (2011, p.12) notes that the third version is related to Schouten's R-indicator and reduces exactly to the R-indicator under certain conditions. We only use the third version for our study, as it was the most interpretable in preliminary analysis with our datasets.

Like the R-indicator, the balance indicator ranges from 0 and 1. When the respondents are "perfectly balanced" (Särndal, 2011, p.8) with the full sample, it equals 1. However, this ideal is not possible to achieve if nonresponse exists (Lundquist and Särndal, 2013, p.23). As the balance between the response set and the full sample decreases, the balance indicator approaches 0. Imbalance stems from two factors: unit nonresponse and differences between respondents and nonrespondents across the auxiliary variables (Särndal, 2011, p.8). Thus, even if the nonresponse rate is relatively low, a large difference between respondents and nonrespondents can create imbalance.

Notation and Formulae:

s = set of sampled units; r = set of respondent units (a subset of s)

$w_i =$ sampling design weight for the i^{th} unit

$P = \frac{\sum_r w_i}{\sum_s w_i}$ = weighted proportion of sampled units that have responded.

$\boldsymbol{D} =$ a vector representing the distance between the response set and the original sample across the subdomains of the auxiliary variables. Each element of the vector is the difference between the weighted proportion of responding units in a particular subdomain versus the weighted proportion of sampled units within that subdomain.

$\boldsymbol{D}' \sum_{\boldsymbol{S}}^{-1} \boldsymbol{D} =$ a single value that represents the lack of balance of the response set with the full sample, where $\boldsymbol{\Sigma_S}$ is a weighting matrix

$B = 1 - 2P\sqrt{\left(\boldsymbol{D}' \sum_{\boldsymbol{S}}^{-1} \boldsymbol{D}\right)}$ = the version of Särndal's balance indicator used in this study.

B measures lack of balance on a 0-1 scale, with 1 occurring only when respondents are in "perfect balance" (Särndal, 2011, p.8) with the full sample.

2.3 Weighted Volume Response Rate

The WVRR is useful for business surveys that estimate totals because it incorporates the tabulation unit's design weight and measure-of-size value (Thompson et al., forthcoming in 2014). Therefore, it places importance on those tabulation units that contribute more to the estimates – i.e., the larger units. Ideally, the measure-of-size variable should be positively correlated with the variable of interest. We define the WVRR as:

$$WVRR = \frac{\sum_{i=1}^{n} w_i \, mos_i \, r_i}{\sum_{i=1}^{n} w_i \, mos_i}$$

Where

n = number of tabulation units for which data collection was attempted

w_i = sampling design weight for the i^{th} tabulation unit

mos_i = measure-of-size variable for the i^{th} tabulation unit

r_i = response indicator variable for the i^{th} tabulation unit

3. Case Studies

In this section, we present the results observed when we retrospectively apply the R-indicator and balance indicator in conjunction with the WVRR at regular intervals over the course of the data period for two business surveys conducted by the U.S. Census Bureau: the *Quarterly Services Survey* (QSS) and the *Business Research and Development and Innovation Survey* (BRDIS), an annual survey. Through our study, we attempt to answer the following questions: do these indicators inform us when the collection strategy has reached phase capacity and if so do our investigative analyses provide insight into which units need to be targeted?

For each survey, we examine four statistical periods of data provided by the survey managers and assess a set of candidate auxiliary variables to find the auxiliary vector that best estimates response and explains the survey variable of interest. To assess the appropriateness of our response propensity model, we perform an *omnibus test* to test the significance of the parameters and the *Adjusted Wald F test* to test whether response is independent of the categories to which we assigned the propensity scores. Additionally, we produce a contingency table to empirically assess how well our model predicts response.

To examine the relationship between the auxiliary variables and the survey variable of interest, we regress the variable of interest against the same set of auxiliary variables. To assess the strength of this relationship, we test the null hypothesis that all regression parameters are equal to zero, similar to the omnibus test above. We use the resulting auxiliary vectors for each program to derive the R-indicator and balance indicator at regular intervals during the data collection period for each survey. We use the resulting values to assess the effectiveness of the collection strategy in obtaining a representative response set.

We make some modifications to reduce complications in our study. The theory for the R-indicator and balance indicator is based on comparing the full sample to the response set (Schouten et al., 2009; Särndal, 2011). In our study, we use only the reporting units for which collection was attempted and not the full set of sampled units. In practice, the WVRR is calculated using tabulation units that derive from the consolidation or split of

reporting units. When the correspondence between survey, reporting, and tabulation units is not one to one, the survey methodologists must appropriately allocate the measure-of-size and the design weight from the sampled survey unit to each reporting unit or tabulation unit. Though the R-indicator and balance indicator are not currently used in practice, a similar adjustment to the design weight should be made if they were used operationally. We do not adjust the design weight or the MOS in any our calculations, including the WVRR and the R-indicator and balance indicator. We do not believe any of our simplifications have a meaningful effect on the findings of our study.

3.1 Quarterly Services Survey (QSS)

The *Quarterly Services Survey* (QSS) is a principal economic indicator series that provides quarterly estimates of revenue and expenses for selected service industries. The QSS is a voluntary survey conducted quarterly with the mail out occurring at the end of each calendar quarter. The QSS sample is comprised of service businesses with paid employees that operate in the covered sectors. The sample design is a stratified design with systematic probability-proportional-to-size sampling. The primary strata are based on industry and tax status. The secondary strata are based on revenue measure-of-size. Sampled units may respond through mail, the internet, fax, or telephone. The sample is updated quarterly to reflect births and deaths. A new QSS sample is selected approximately every five or six years.

References:

- http://www.census.gov/services/qss/qsstechdoc.html
- http://www.census.gov/econ/overview/se0600.html
- U.S. Census Bureau (2014)
- Weidenhamer and Ferreira (2014)

For our retrospective study, we employed QSS data from four quarters: 2013 Q2, 2013 Q1, 2012 Q4, and 2012 Q3. We limited our analysis to the key estimate of interest, *quarterly estimates of revenue* (QREV) and to those reporting units for which data collection was attempted. Analysis was conducted at the six-digit industry level. Each industry had a "certainty stratum" (i.e., reporting units that have a design weight of one) and a varying number of noncertainty strata.

3.1.1 Selection of the QSS Auxiliary Vector

The R-indicator and balance indicator both assess the degree to which the respondents are similar to the full sample in terms of an auxiliary vector that should be both predictive of response and is correlated with the estimate(s) of interest, *QREV*. For QSS, we found that the auxiliary variable "stratum[2]," a function of the measure-of-size[2] variable best satisfies these two criteria.

The fitted logistic regression model is shown below:

$$\text{logit}(\hat{\pi}_{i,j,z}) = \alpha_z + \beta_{j,z} X_{i,j,z}$$

$\hat{\pi}_{i,j,z}$ = estimated response propensity; X = auxiliary vector;
i = reporting unit; j = stratum; z = industry

[2] Stratum (STRATM) is a variable in StEPS that for QSS represents the secondary strata which are based on revenue measure-of-size. The measure-of-size variable is typically either census receipts or administrative receipts that have been inflated/deflated using administrative payroll to put it in current year terms.

The omnibus test in Table 1 below indicates that the overall fit of this model was significant in nearly all industries for all quarters. However, these results do not prove that our auxiliary variables are good predictors of response, which is why we also examine the distribution of response across the propensity categories (see Table 2).

Table 1. Fit of Response Propensity Model for QSS

Period	*Omnibus Test**
2013 Q2	99.3%
2013 Q1	97.8%
2012 Q4	99.3%
2012 Q3	99.3%

*Percent of Industries where model is significant (α = 0.10). {Total number of industries: 142 for all periods}

From conducting an Adjusted Wald F test, we reject the null hypothesis that response is independent of the propensity categories for each period (α = 0.10). This means there is some relationship between response and the propensity categories shown, but the relationship is not strong. For the 2013 Q2 sample, the model designates 51.5 percent of the respondents as having either a "medium" or "high" response propensity, but the model also assigns medium or high propensities to 29.2 percent of the nonrespondents (see Table 2). Similar results occur in the prior quarters. This is a mixed bag. The model does a fair job of classifying respondents. However, a better model would more clearly distinguish between respondents and nonrespondents. Having assessed our model's ability to predict response, we now turn our attention to its ability to explain the study variable, QREV.

Table 2. Weighted Response Propensity Proportions for QSS

Period	*Response Status*	*Very Low* $\hat{\rho} \leq .50$	*Low* $.50 < \hat{\rho} \leq .60$	*Medium* $.60 < \hat{\rho} \leq .70$	*High* $.70 < \hat{\rho}$
2013 Q2	R	22.4	26.1	33.0	18.5
	NR	43.9	26.9	23.4	5.8
2013 Q1	R	29.6	32.5	22.5	15.3
	NR	51.5	30.0	14.5	4.1
2012 Q4	R	27.9	34.2	22.9	15.0
	NR	49.5	31.5	14.6	4.4
2012 Q3	R	26.9	35.3	22.3	15.5
	NR	48.2	33.0	14.0	4.7

R = respondent; NR = nonrespondent

We use the below model to assess how well QREV correlates with our auxiliary vector:

$QREV_{i,j,z} = \alpha_z + \beta_{j,z} X_{i,j,z} + \varepsilon_{i,j,z}$
X = auxiliary vector; ε = error term;
i = reporting unit; j = stratum; z = industry

Table 3 shows that model is significant in explaining the key variable, QREV in all industries for all quarters studied, demonstrating that our auxiliary vector is strongly related to revenue. Since our auxiliary vector is also adequate in estimating response, we proceed to analyze the survey returns via the R-indicator and balance indicator.

Table 3. Fit of Study Variable Model for QSS

Period	*Omnibus Test**
2013 Q2	100%
2013 Q1	100%
2012 Q4	100%
2012 Q3	100%

*Percent of Industries where model is significant (α = 0.10). {Total number of industries: 142 for all periods}

3.1.2 Analyzing the Composition of Survey Response for QSS During Data Collection

In Figure 1 below, we use the R-indicator and balance indicator to assess the representativeness of the response set during data collection for the second quarter of 2013 for QSS (results are similar for the other periods and are not shown). For the data collection period, we divide the final proxy unit response rate (URR) of 0.65 into 10 intervals and calculate the indicator values at the end of each interval. However, since the size of the responding units is important in business surveys and because the R-indicator and balance indicator tend to register high values at the beginning of data collection due to the small number of respondents, we use the weighted volume response rate (WVRR) to provide a signal for when a sufficient number of returns have been received. For both surveys, we decide on an arbitrary WVRR value of 0.25 to be an appropriate starting point (beginning of "analysis zone"), as at least one quarter of the sampled population is accounted for. [Note: In Figure 1, the fact that the WVRR value of 0.25 is approximately equal to the proxy URR value of 0.26 is a coincidence.] At and beyond this point, we obtain more meaningful values for the response indicators.

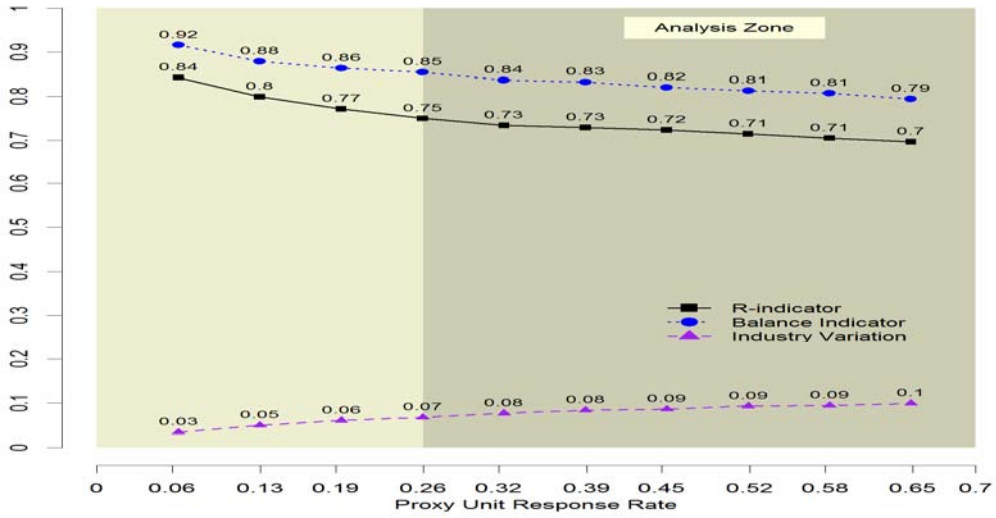

Figure 1: Assessing Response Representativeness across Industries for QSS 2013 Q2

We begin our analysis in the "analysis zone," where the R-indicator and balance indicator attain initial values of approximately 0.75 and 0.85, respectively, and then decline in value throughout the remainder of the collection period. Thus, although the response set increases in size, its representativeness is declining. One possible explanation is the differential proxy URRs amongst the industries sampled, as indicated by the purple line in Figure 1, which provides a measure of the variation in industry response throughout the collection period. In the "analysis zone," this variation increases from a low of 0.07 to a high of 0.10. Though this increase is rather small, the fact that it is increasing shows

that the additional respondents are not improving the representativeness of the response set. In fact, they only increase the disparity in response rates between industries. We also examine the variation in response across strata (Figure 2) and find a similar result. As with the industry level examination, while the actual increase in variation across strata is relatively small, the fact that stratum-level variation is increasing at all suggests that increasing response does not increase response representativeness across strata.

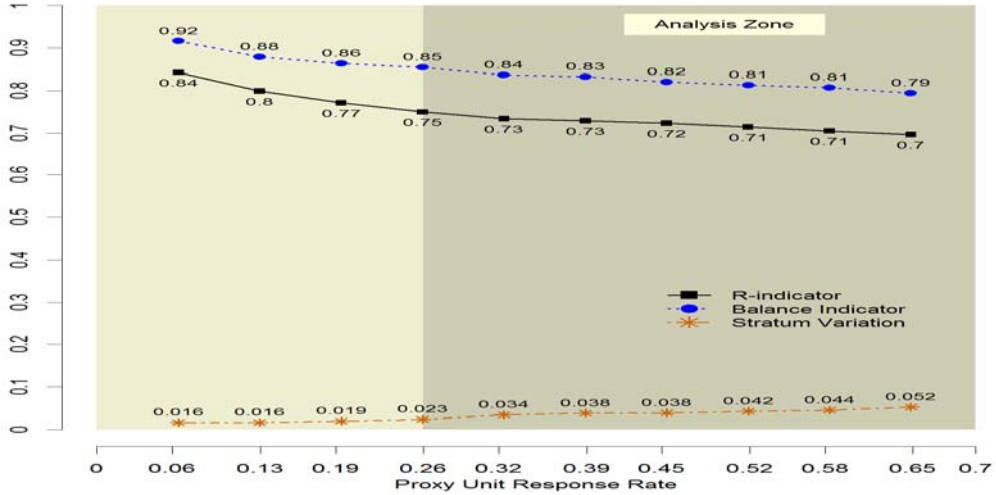

Figure 2: Assessing Response Representativeness across Stratum for QSS 2013 Q2

Given that QSS analyst review and contact procedures focus on obtaining responses from the larger units, we compare the response over time by size of unit using certainty status (certainty versus noncertainty units) as a proxy for unit size (see Figure 3). Though there is some differential response between large and small units, it is rather small (from a low of 0.04 to a high of 0.07 in the "analysis zone"). Initially, the small units have a higher response rate than the larger units. However, as the data collection period draws to an end, the analysts make a concerted effort to obtain response from these important units; thus, the large units respond at a higher rate by the close of the collection period. This information confirms our earlier suspicions that the collection procedures implemented do not obtain a representative sample, and the decline in "representativity" over the collection period is attributable to the focus on obtaining response from the larger units.

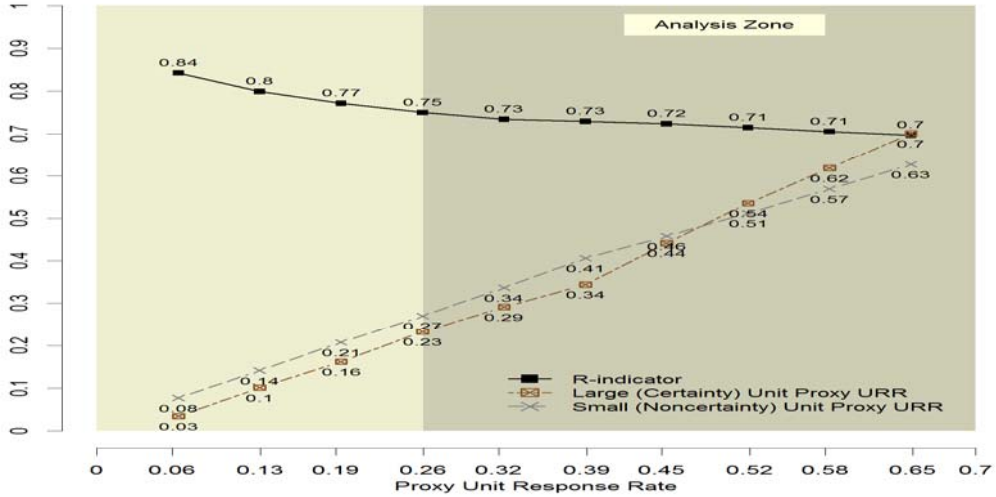

Figure 3: Assessing Response Representativeness across Size of Unit for QSS 2013 Q2

3.1.3 Summary of Response Set Analysis for QSS

In summary, for QSS we find that the R-indicator and balance indicator both declined throughout data collection, indicating that additional response only reduced the representativeness of the response set. This result is not a surprise, as the subject-matter experts from QSS informed us that their collection strategy prioritizes obtaining response from the largest units. Although this strategy may reduce variance in the resulting estimates, it may also have consequences in terms of nonresponse bias. Because there is a concentration of certainty units in a small proportion of industries, this strategy may help explain the increasing variation in response across industries, as seen in Figure 1. As a result, the potential for nonresponse bias may be higher in industries with small numbers of certainty units that receive less attention for nonresponse follow-up. Though pinpointing exactly where the current strategy reached its "phase capacity" is difficult, if not impossible, it is clear that the current strategy is not obtaining a representative sample. Thus, if the survey managers aim to reduce the risk of nonresponse bias, they should target subdomains that are lagging in response.

3.2 Business Research and Development and Innovation Survey (BRDIS)

The *Business Research and Development and Innovation Survey* (BRDIS) provides primary source data pertaining to research and development (R&D) performed or funded by business within the United States (Hough 2011, p.3). The U.S. Census Bureau conducts BRDIS as part of an interagency agreement with the survey's sponsor, the National Center for Science and Engineering Statistics (part of the National Science Foundation) (Hough 2011, p.4). The BRDIS is conducted annually with the collection period typically beginning in March or April and ending in the latter part of December. The BRDIS sample is comprised of roughly "43,000 companies with locations in the United States that are classified in select manufacturing and nonmanufacturing industries" (Hough 2011, p.3). The sample design is a stratified design that uses either simple random sampling or PPS sampling within strata to select units. Stratification is based on R&D activity and a NAICS-based industry code. The major strata are known positive R&D activity, unknown R&D activity, and known zero R&D activity. Unlike QSS, response to BRDIS is mandatory by law. Companies known to perform large amounts of R&D and companies with large amounts of payroll are selected with certainty. For more details concerning BRDIS, please see http://www.nsf.gov/statistics/srvyindustry/.

3.2.1 Selection of the BRDIS Auxiliary Vector

For BRDIS, we limited our analysis to units on the sampling frame that were in the known positive R&D activity stratum as this where most of the contribution to the R&D estimates comes from. Once again, we fit logistic and linear regression to candidate variables provided by subject-matter experts to find the set that best explains both response and the study variable, *R&D expenditures* (RDEXP), which represents the worldwide R&D expenditures of U.S. firms.

We first give background on an auxiliary variable that is very relevant to how the survey returns are received: form type. There are two form types. Units with more than one million in R&D expenditures in at least one prior period dating back to 2007 were sent the long form in 2012. In 2012, the long form[3] was 48 pages and estimated to take an average of 14.3 hours to complete; all other units received a short screener form[4] (short

[3] http://www.nsf.gov/statistics/srvyindustry/about/brdis/surveys/srvybrdis_2012_BRDI-1.pdf
[4] http://www.nsf.gov/statistics/srvyindustry/about/brdis/surveys/srvybrdis_2012_BRD-1S.pdf

form) which was only 8 pages long in 2012 and estimated to take an average of 1.5 hours to complete. Because the long form is much more burdensome to complete than the short form, units receiving the long form often do not complete it until later in the collection period while many units receiving the short form that no longer conduct R&D return it immediately. Our selected model, which incorporates the size of the reporting unit (certainty versus noncertainty); the form type used (long form versus short form); and a unit's measure-of-size (previous R&D value) is shown below:

$$\text{logit}(\hat{\pi}_{i,j,k,l,z}) = \alpha_z + \beta_{j,z}X_{i,j,z} + \beta_{k,z}X_{i,k,z} + \beta_{l,z}X_{i,l,z}$$

$\hat{\pi}_{i,j,k,l,z}$ = estimated response propensity; X = auxiliary vector;
i = reporting unit; j = size of unit; k = form type; l= measure-of-size; z = industry

We also examined interaction effects between auxiliary variables, but including interactions did not meaningfully improve our analysis, hence we use the above model as it is more parsimonious. Table 4 below indicates that the overall model is significant for roughly two-thirds of the industries examined in 2012. Results vary by year, but for all years there is a noticeable number of industries for which the overall model is not significant. Thus, our model is appropriate for some of the industries, but not all of them.

Table 4. Fit of Response Propensity Model for BRDIS

Period	*Omnibus Test**
2012	66.7%
2011	54.1%
2010	76.7%
2009	49.2%

*Percent of industries where model is significant (α = 0.10). {Total number of industries by period: 2012 (60); 2011 (61); 2010 (60); 2009 (65)}

To better examine how well our model distinguishes respondents from nonrespondents, we compare the weighted proportion of units classified in each score category in Table 5 as we did with the QSS data in Table 2. From conducting the Adjusted Wald F test, we conclude that response is not independent of the propensity categories for each period (α = 0.10). However, this finding means little in practical terms as we see when analyzing the distribution of response propensities by response (Table 5). For all periods, the model assigns a "medium" or "high" response propensity value to nearly all units, regardless of whether they responded or not. Thus, our propensity model is not a good fit; it is blindly assigning large propensities to all units and not distinguishing between respondents and nonrespondents with any degree of success. In a related study, the branch chief of BRDIS had similar difficulty estimating response propensities (Hough and Shackelford, forthcoming in 2014).

It is possible that response is related to a latent class variable that is not captured adequately by the auxiliary data. It is also possible that the nonresponse mechanism is not missing-at-random (NMAR) for these cases, as they are in the "known" R&D stratum, and units are not responding for reasons directly related to the study variable, RDEXP. If nonresponse is directly related to the study variable, RDEXP, for reasons not captured by our auxiliary vector, then the logistic regression model propensities will be inadequate predictors and will not provide useful subdomains for investigating causality.

Table 5. Weighted Response Propensity Proportions for BRDIS

Period	*Response Status*	*Very Low* $\hat{\rho} \leq .50$	*Low* $.50 < \hat{\rho} \leq .60$	*Medium* $.60 < \hat{\rho} \leq .70$	*High* $.70 < \hat{\rho}$
2012	R	0.9	1.5	10.3	87.3
	NR	5.5	4.2	18.1	72.2
2011	R	0.2	0.9	14.8	84.1
	NR	1.1	2.4	24.0	72.5
2010	R	0.9	2.0	13.0	84.2
	NR	5.1	4.9	21.0	68.9
2009	R	0.2	0.9	6.7	92.2
	NR	1.6	2.9	14.2	81.3

R = respondent; NR = nonrespondent

Although our response propensity model is extremely weak, the response indicators may be informative for assessing the composition of the response sample if the variables included are related to the variable of interest. To assess how well these auxiliary variables explain RDEXP, we fit the following regression model:

$RDEXP_{i,j,k,l,z} = \alpha_z + \beta_{j,z}X_{i,j,z} + \beta_{k,z}X_{i,k,z} + \beta_{l,z}X_{i,l,z} + \varepsilon_{i,j,k,l,z}$
X = auxiliary vector; ε = error term;
i = reporting unit; j = size of unit; k = form type; l = measure-of-size; z = industry

Table 6 shows that model is significant in explaining the key variable, RDEXP, in all industries for all periods studied, demonstrating that our auxiliary vector is strongly related to worldwide R&D expenditures. However, as previously shown, our auxiliary vector is not useful for estimating response. We still proceed to analyze the survey returns via the R-indicator and balance indicator, but we must temper our findings by the fact there is an unknown response mechanism not captured by our auxiliary vector.

Table 6. Fit of Study Variable Model for BRDIS

Period	*Omnibus Test**
2012	100%
2011	100%
2010	100%
2009	100%

*Percent of industries where model is significant ($\alpha = 0.10$). {Total number of industries by period: 2012 (61); 2011 (61); 2010 (61); 2009 (67)}

3.2.2 Analyzing the Composition of Survey Response for BRDIS During Data Collection
In Figure 4, we use the R-indicator and balance indicator to assess the representativeness of the response set throughout the collection period for 2012 BRDIS. We again use a WVRR of 0.25 as a starting point for our analysis. In the analysis zone, the R-indicator and balance indicator steadily increase while the variation of response by industry declines, suggesting that additional response increases the representativeness of the resulting response set. Thus, we cannot conclude that the collection strategy used reached a phase capacity during data collection as there is no indication that the current strategy of using the WVRR by the three major strata to monitor priority industry categories (based on R&D intensity) needs to be altered to improve representativeness across

industries. However, our indicators are only as good as the auxiliary variables, and for BRDIS we are missing important covariates to better model the response mechanism.

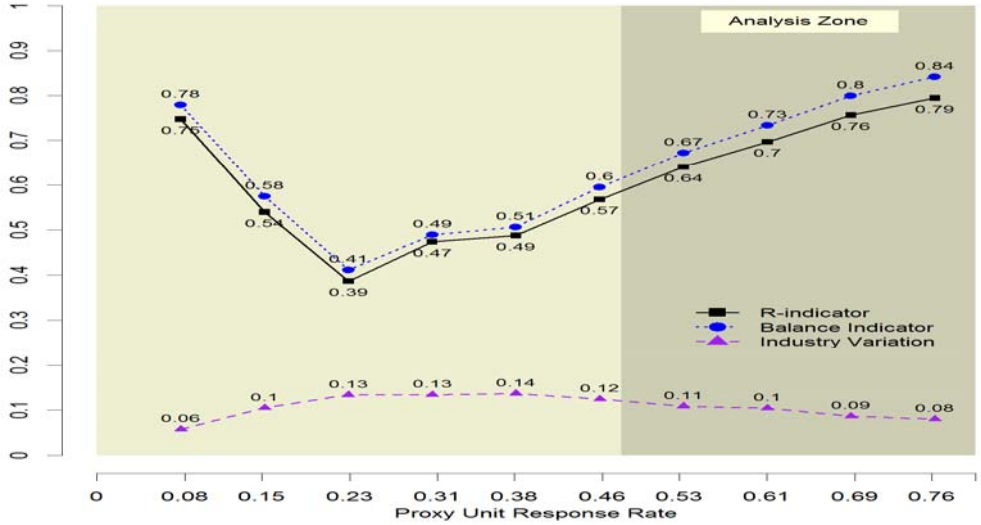

Figure 4: Assessing Response Representativeness across Industries for BRDIS 2012

In Figure 5, we examine the survey response over time as a function of form type. Initially, the units that respond by the short form respond at a larger rate than those that complete the long form. However, towards the end of the collection period this gap narrows. Though not shown, we also found that there is initially a large differential response between the small (noncertainty) and large (certainty) units, but as the collection period came to a close this gap decreased as the large units began to respond at a higher rate than the small units.

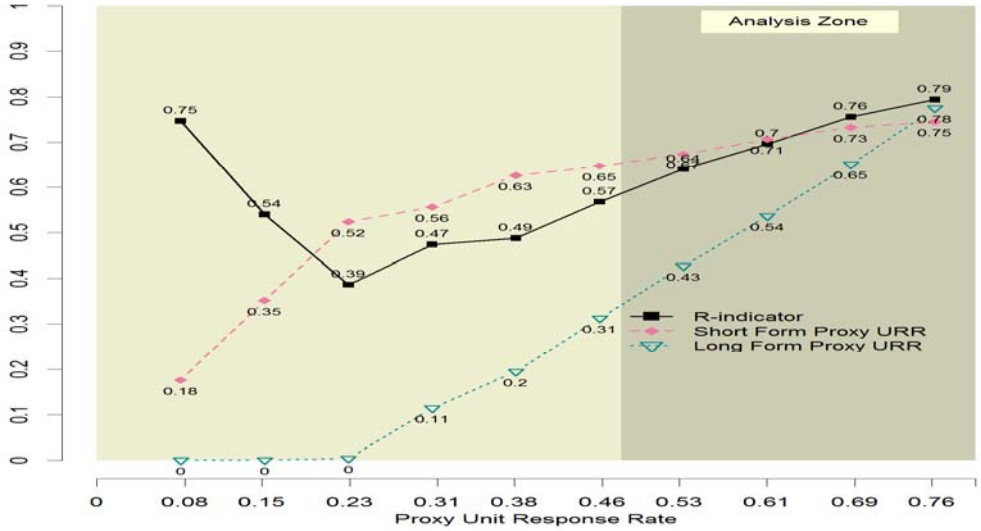

Figure 5: Assessing Response Representativeness across Form Type for BRDIS 2012

3.2.3 Summary of Response Set Analysis for BRDIS

In summary, for BRDIS the response indicators show that as additional response was received, the collection strategy employed produced a more representative response set across industry, form type, and unit size. However, these results must be tempered by the fact that our model is not capturing the full response mechanism. The BRDIS subject-

matter expert surmises that nonresponse is tied to the collection instrument (i.e. the length and detail of the form). Hence, the missingness is not random. If the true response mechanism is not missing-at-random and units are not responding for a reason directly related to their R&D expenditures (for example, the nonrespondents have considerable R&D but view the form as burdensome), then the response indicators are painting an incomplete picture as not only is the response set unrepresentative, but there will be considerable non-response bias in the resulting estimates.

4. Conclusion

The purpose of this research was to investigate the potential use of the R-indicator and balance indicator in conjunction with the weighted volume response rate (WVRR) during data collection for assessing the effectiveness of the collection strategy in obtaining a representative response set. We applied these indicators jointly to two business surveys: the Quarterly Services Survey (QSS) and the Business Research and Development and Innovation Survey (BRDIS). We found that challenges faced by business surveys, such as limited auxiliary information related to response and the importance of obtaining response from large units limited their applicability for business surveys.

For both business surveys, it was difficult to find auxiliary information that was related to the response mechanism. As a result, the response indicators did not provide a complete picture of the composition of the response set. The upward trend seen in these indicators during the BRDIS collection period is only indicative of increasing representativeness of response among the specific variables examined. The effect of the true, but unknown, response mechanism on response composition could not be assessed by the indicators with the available auxiliary information. Thus, the indicators may not be appropriate for informing the BRDIS survey managers when, or if, to alter their collection strategy.

Targeting the larger units makes sense for business surveys. Doing so will reduce the variance of the estimate but may increase the risk of nonresponse bias. Thus, for surveys that employ this collection strategy, the R-indicator and balance indicator are not as useful. A more informative measure is the WVRR. BRDIS already relies on this measure to track its prioritized industry categories, which are based on R&D intensity.

Despite these limitations, the R-indicator and balance indicator were effective in analyzing the representativeness of the response set, with respect to the specific auxiliary information included. For both programs, the trends in the indicators were consistent with the variation in response across subdomains of the auxiliary variables. If survey managers account for the indicators' shortcomings when assessing representativeness, then these indicators may be informative tools for helping them assess their collection strategies.

Acknowledgements

The authors thank Richard Hough, David Kinyon, Harold Laney, Jim Liu, Mary Mulry, Aidan Smith, and Jenny Thompson for their review and suggestions. The authors also owe a special thanks to Hannah Thaw for creating the figures used throughout this report.

References

Groves, R.M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, *70* (5), 646-675.

Groves, R.M. and Heeringa, S.G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A*, *169* (Part 3), 439-457.

Hough, R. (2011). 2011 Business R&D and Innovation Survey (BRDIS) Methodology Report. *Internal Memorandum.* Research, Development, and Innovation Surveys Branch; U.S. Census Bureau; Washington, DC. Available upon request.

Hough, R. and Shackelford, B. (forthcoming in 2014). An Examination of Nonresponse in the Business R&D and Innovation Survey. *In JSM Proceedings, Government Statistics Section.* Alexandria, VA: American Statistical Association.

Kreuter, F. (Ed.). (2013). *Improving Surveys with Paradata: Analytic Uses of Process Information*. Hoboken, NJ: John Wiley & Sons, Inc.

Little, R.J. and Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology*, *31 (2)*, pp. 161-168.

Lohr, S. (1999). *Sampling: Design and Analysis*. Pacific Grove, California: Brooks/Cole Publishing Company.

Lundquist, P. and Särndal, C-E. (2013). *Responsive design, Phase II – Features of the nonresponse and applications.* Statistics Sweden, Research and Development – Methodology Reports from Statistics Sweden.

Oliver, B. and Thompson, K.J. (2013). An Analysis of the Mixed Collection Modes for Business Surveys at the U.S. Census Bureau. *Proceedings of the Federal Committee on Statistical Methodology*, pp. 1-15.

Peytcheva, E. and Groves, R.M. (2009). Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates. *Journal of Official Statistics*, *25 (2)*, 193-201.

Särndal, C-E. (2011). The 2010 Morris Hansen Lecture Dealing with Survey Nonresponse in Data Collection, in Estimation. *Journal of Official Statistics*, *27 (1)*, pp. 1-21.

Schouten, B., Cobben, F., and Bethlehem, J. (2009). Indicators for the Representativeness of Survey Response. *Survey Methodology*, *35 (1)*, pp. 101-113.

Schouten, B., Shlomo, N., and Skinner, C. (2011). Indicators for Monitoring and Improving Representativeness of Response. *Journal of Official Statistics*, *27 (2)*, pp. 231-253.

Thompson, K.J. and Oliver, B. (2012). Response Rates in Business Surveys: Going Beyond the Usual Performance Measure. *Journal of Official Statistics*, *28 (2)*, pp. 221-237.

Thompson, K., Oliver, B., and González, Y. (forthcoming in 2014). Response Rates as Process Control Tools: Creative Usage of Existing Performance Metrics. *In JSM Proceedings, Government Statistics Section.* Alexandria, VA: American Statistical Association.

U.S. Census Bureau (2014). Sampling Methodology Inventory for Non-Reimbursable Programs in the Economic Directorate of the United States Census Bureau. *Internal Memorandum (EDMS Document #224451).* U.S. Census Bureau; Washington, DC. Available upon request.

Wagner, J. (2012). Research Synthesis—A Comparison of Alternative Indicators for the Risk of Nonresponse Bias. *Public Opinion Quarterly*, *76* (3), pp. 555-575.

Weidenhamer, D. and Ferreira, L. (2014). What is the Economic Area Sampling Methodology Inventory? (Part 3 of 3). *Internal PowerPoint presentation presented at the Economic Area Methodology Seminar.* Program Research and Development Branch; U.S. Census Bureau; Washington, DC. March 19, 2014. Available upon request.