Conducting Nonresponse Bias Analysis for Business Surveys

Joanna Fane Lineback and Katherine J. Thompson
Office of Statistical Methods and Research for Economic Programs
U.S. Census Bureau[1], Washington, D.C. 20233

Abstract
The Office of Management and Budget directs survey programs to conduct nonresponse bias analysis when response rates fail to meet target values. The literature focuses largely on nonresponse bias analysis methods for demographic surveys. Such surveys are generally characterized by multi-stage designs with heterogeneous populations within selected clusters. In contrast, business surveys are characterized by single-stage designs with highly skewed populations. This paper examines nonresponse bias analysis methods for business surveys, including response rate analysis, the use of frame data, and the examination of the response prediction and propensity models, illustrating each method with examples from ongoing economic programs conducted by the U.S. Census Bureau.

Key Words: nonresponse bias analysis, response rate, business survey, response model, response mechanism

1. Introduction

Perhaps the biggest issue in survey research is the problem of missing data. In the missingness is the potential for nonresponse bias. Bias in survey estimates could lead to incorrect conclusions about the population of interest. Although low response rates do not necessarily indicate nonresponse bias (Groves et al., 2008), they are often used as a measure of data quality.

The Office of Management and Budget (OMB) directs survey programs to make plans to evaluate nonresponse bias when expected unit response rates are less than 80% (Graham, 2006). Some survey programs have even more stringent in-house policies. While the 80% cutoff is somewhat arbitrary, this directive is a start for survey programs to address nonresponse bias. Much of the research in this area focuses on conducting nonresponse bias studies for household surveys. We need additional tools for analyzing business surveys' nonresponse.

This paper is meant to serve as a practical tool for researchers conducting nonresponse bias studies for business or establishment surveys. It is also meant to generate discussion in this area of survey research, as there is still much work to do.

The remainder of this paper is divided into four sections. The next section addresses the unique properties of business surveys. The third section presents commonly used methods of assessing nonresponse bias, focusing on methods particularly helpful in the

---

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

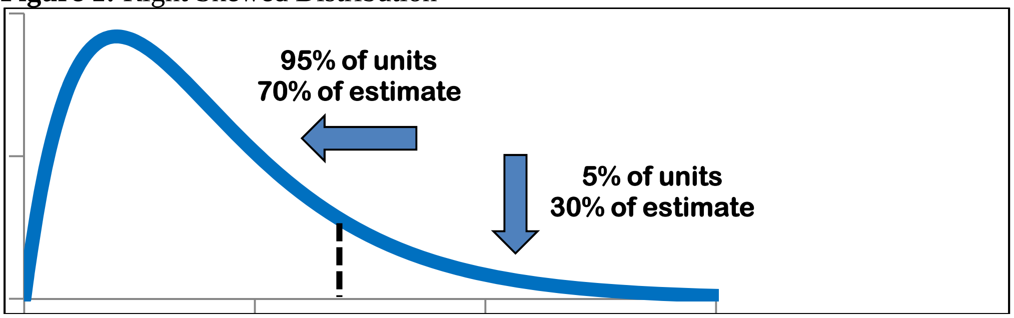
world of business surveys. Section 4 provides illustrative examples of the methods presented in Section 3, with selected analyses from three cases studies conducted at the U.S. Census Bureau. The conclusion provides nonresponse bias questions to consider for future research.

2. Characteristics of Business Surveys

2.1 Certainty Status

Business data tend to follow a skewed distribution (e.g., sales and inventories). Consider the following graph presenting a fictional business data distribution.

Figure 1: Right Skewed Distribution



Businesses with values at the high end of the scale are often those whose value legitimately comprises the majority of the tabulated total.[2] These units are often the ones of greatest interest. To ensure that the survey sample is representative of the population, these units are often included in the sample "with certainty," meaning they have a design weight of one and every effort is made to obtain their data.

2.2 Reporting Versus Tabulation Units

Business surveys may need to distinguish between the "reporting unit" and the "tabulation unit." A reporting unit is one that has been established for the purpose of collecting survey data. A tabulation unit is one used for estimation. For example, company (reporting unit) data might be divided into several distinct tabulation units – each representing an industry – for analysis and estimation. Alternatively, a company might request to *report* data by geographic location, and these data would need to be combined back to the company level for tabulation purposes. For many survey programs, there is no distinction between reporting and tabulation units.

2.3 Response Rates

The unit response rate (URR) formula employed by economic surveys and censuses at the U.S. Census Bureau is

$$URR = [R/(E+U)] \times 100, \qquad (1)$$

where
R is the count of *reporting units* selected for the sample that were eligible for data collection and classified as a response,

---

[2] For example, consider the grocery store industry, where the majority of business is conducted by large retail chains, not "mom and pop" establishments.

E is the count of reporting units selected for the sample that were eligible for data collection, and
U is the count of reporting units selected for the sample for which eligibility could not be determined (Bates et al., 2008).

Bates et al. (2008) define a respondent as an eligible reporting unit for which an attempt was made to collect data, the unit belongs to the target population, and *the unit provided sufficient data to be classified as a response.* To satisfy the latter requirement, each program determines which collected data items are *required* in advance of data collection. Response status for each unit is determined after all data processing – including analyst review/corrections and machine editing/imputation – is completed. As the data quality restrictions (i.e., edits) on required data items for the program increase, the greater the likelihood of the URR decreasing, because it is more difficult for reporting units returning the questionnaire to qualify as a respondent. Consequently, it is possible to offer specific protocols designed to improve the amount of contact with nonrespondents that do *not* improve the URR. For example, a program might require follow-up telephone interviews with a contact person unable to provide the requested data – commonly occurring when the company contact operates in a non-accounting office.

Because business data tend to be highly skewed, sampled cases with large design weights often contribute very little to overall tabulated totals. To avoid over-representation of such small cases in computation, URRs are computed without using design weights.

For evaluating a business program's nonresponse bias potential, the above rate computed at the program level can lead to erroneous analysis. In a business survey setting, it is more telling to calculate and analyze URR by subgroup, usually starting with certainty status. If certainty status is an indicator of importance in the sense that the data from these units is necessary to make inference about the population, then we are particularly interested in the response rate for these units. Moreover, since we are taking a census of these "important" units, any nonresponse causes nonresponse bias. It is unlikely that this source of nonresponse bias can be corrected via modeling (e.g., weighting or imputation), since certainty units by definition are unique. Consequently, the URR can be an inconsistent measure of data quality and a poor predictor of nonresponse bias.

In economic surveys, the more consistent measures of data quality include a measure of size (MOS) (e.g., payroll, capital expenditures) to account for the unit's relative importance in the estimates (Tucker et al., 2007). At the U.S. Census Bureau, economic programs compute Total Quantity Response Rates (TQRRs) for *each key* data item as

$$TQRR = \left[\frac{\sum_{i=1}^{N_T} w_i \times (r_{ti} + q_{ti}) \times |t_i|}{\sum_{i=1}^{N_T} w_i f_i |t_i|}\right] \times 100, \qquad (2)$$

where
w_i is the design weight of tabulation unit *i*,
r_{ti} is the indicator variable for reported data for tabulation unit *i* and data item *t*,
q_{ti} is the indicator variable of "equivalent quality" data[3] for tabulation unit *i* and data item *t*,

[3] Equivalent quality data are *indirectly* received from the "respondent." To be considered equivalent quality, substituted data *should* be validated by an independent report.

t_i is the data value for unit *i*,
f_i is the nonresponse weighting adjustment factor for tabulation unit *i*, and
N_T is the total number of eligible tabulation units.

In other words, this is the proportion of the weighted total of data item *t* obtained from directly reported and "equivalent quality" data (Bates et al., 2008).[4] Note that the denominator includes any imputed values, and, depending on the survey, there may not be a nonresponse weighting adjustment factor.

The TQRR for a given item can be very different than the program's URR. For example, the TQRR for sales might be 90%, when the overall URR is only 70%.

In-house procedures for analyst review and follow-up of survey data are designed to improve the quality of the *estimates*. Analysts strive to reduce imputation rates for *all* key items (imputation rate for a given item = 1 – TQRR). This is usually best accomplished by nonresponse follow-up of large cases expected to contribute substantially to the estimate, followed by intensive analyst research for "large impute" cases comprised of more phone calls (targeted questions) and searches for auxiliary data sources (e.g., financial reports) to replace imputed values with equivalent data. For programs that publish information on one or two characteristics, the TQRRs for each item are clearly superior performance measures over URRs. However, if the survey publishes several key characteristics, it may be unwise to measure performance by setting target TQRR values for all characteristics, since meeting all target values may be difficult or even impossible depending on the number of collected items and the processing cycle time allotment.

2.4 Methods of Handling Nonresponse

Ideally, valid data are received from every eligible sampled unit; however, this is never the case. Like household surveys, business surveys often use one of the following methods to account for nonresponse: follow-up, imputation, or weighting adjustments.

For our purposes, nonresponse follow-up refers to the attempt to convert nonrespondents to respondents. Depending on the survey's sample size and timing, this is often reserved for certainty or large noncertainty units. For example, a monthly survey may restrict follow-up exclusively to large cases, whereas an annual survey may perform most of its follow-up on large cases, but may follow-up on smaller cases later in the processing cycle. Weighting adjustments are done at the unit level and involve some design weight adjustment such as a ratio adjustment (very common for business surveys) or raking to know totals (not so common). Imputation is done at the unit or item level and is the process of creating non-missingness by inferring from other data what a missing value "should" be (Singh and Petroni, 1988). When administrative data are available for all key items, direct substitution is generally preferred to model imputation: this method is in fact used on "small reporters" who are not sent questionnaires by design for selected programs (to reduce respondent burden). Following direct substitution, the most commonly used imputation methods in the Census Bureau's Economic Directorate are variations of regression models, such as ratio imputation, although a few programs employ mean or median imputation models or hot deck. At the U.S. Census Bureau, eight economic programs use nonresponse weighting adjustments, and 48 economic

[4] If no auxiliary data are available, then it might be useful to calculate the quantity response rate. This is essentially the TQRR without equivalent quality data.

programs use imputation (Ozcoskun and Hayes, 2009). Most programs perform nonresponse follow-up, at least for certainty cases.

All three adjustment methods make assumptions about the data's response mechanism. Follow-up assumes the response mechanism generating missing data is non-ignorable (i.e., that respondents and nonrespondents values are systematically different in the collected data items). Imputation and weighting adjustments assume an *ignorable* response mechanism such as a missing at random (MAR) or covariate-dependent mechanism (Little and Rubin, 2002 and Shao and Thompson, 2009). That is, they assume that responses differ systematically based on covariates, but that these covariates are not directly related to the items under consideration. Thus, the respondents comprise a random subsample, and inference can be conducted after employing an adjustment model – such as imputation or a weighting adjustment – to the respondent data. Most of the examples cited below assume an MAR mechanism, where nonresponse is uniform within weighting or imputation cells.

In assessing the impact of nonresponse bias, it is important to validate the response model assumptions that justify the nonresponse adjustment method. Otherwise, these methods can add bias, creating the need to reevaluate imputation and weighting methods.

2.5 Frame Data

Frame data are usually available for respondents and nonrespondents. Business sampling frames often include a MOS variable which is highly correlated with key data items. For example, MOS may be the value of sales reported by the unit in the Economic Census, and the key item for the survey may be annual or monthly sales. If the frame data *are* highly correlated with one or more key data items collected in a survey, systematic comparisons between respondents and nonrespondents can be made.

3. Methods for Investigating Nonresponse Bias

3.1 Getting Started

Before starting a nonreponse bias study, at a minimum, it is helpful to ask the following questions:

- What nonresponse bias analysis has been done for this survey program before?
- What is the sample design?
- What are the modes of data collection?
- What are the variables of interest?
- How is nonresponse currently handled?
- How is estimation performed?
- Are there known areas (e.g., industries) to target in nonresponse bias analysis?
- What are the publication cells?
- Are there any reporting-tabulation unit issues that might affect inference?
- How many statistical periods are of interest?
- Are there useful frame data available?

3.2 Methods Used to Study Nonresponse Bias

3.2.1 Response Rates

One method of studying nonresponse bias is to examine response rates. For business surveys, it is helpful to examine response rates by subgroup, including certainty status or other characteristics that may be building blocks in the survey's sample design. An example using response rates to identify *potential* nonresponse bias for the Monthly Retail Trade Survey is provided in Section 4.2.

Unit and Total Quantity Response Rates

Equation (1) is used to calculate URRs for business surveys at the U.S. Census Bureau. Low URRs in specific subgroups might indicate somewhere to look for potential nonresponse bias. Like URRs, TQRRs [equation (2)] can be helpful when examined on subgroups. Low TQRRs in select subgroups are a red flag for potential nonresponse bias.

Response Rates Over Time

One low response rate is not necessarily indicative of the trend for a given survey program. For instance, for a monthly survey, one month out of a year could have low response relative to other months. In this case, it is worthwhile to determine whether the low response rate is a one-time event or a seasonal phenomenon. Response rates over time – both URR and TQRR – should be considered.

3.2.2 Compare Nonrespondents and Respondents on Frame Variables

As discussed in subsection 2.5, a method for finding potential nonresponse bias is to compare respondents and nonrespondents across a frame variable that is highly correlated[5] with a variable of interest (Harris-Kojetin, 2009). By definition, frame data are available for both respondents and nonrespondents. Assuming the missingness within imputation cells is ignorable, the researcher could conduct an experiment using frame data to test the null hypothesis that there is no difference between estimates for respondents and nonrespondents. Parametric statistical tests (e.g., t-test) could be helpful if we can make large sample assumptions. Otherwise, nonparametric statistical tests to assess distributional differences based on an auxiliary variable (e.g., Wilcoxon tests, Pearson's chi-square test) are an option. If the response mechanism cannot be validated, then the adjustment method used could induce a model bias without ameliorating the nonresponse bias (Thompson, 2009). In Section 4.3, results from a nonresponse bias study for the Quarterly Services Survey exemplify this method.

3.2.3 Response Propensity/Prediction

To minimize bias, we should look into the validity of the response propensity and prediction models used to adjust respondent data. A prediction model predicts a value of a variable y. The response propensity model relates covariates to an individual unit's probability of response. Under an MAR response mechanism, if the covariate used to develop the weighting and imputation cells is highly correlated with the probability of response and the units within the weighting and imputation cells have the same cell mean, then nonresponse bias is minimized by weighting the respondent data with the inverse response rate or imputing missing observations with the cell mean value. If the model does not hold then we could introduce bias.

Similarly, if the weighting or imputation method used to account for unit nonresponse explicitly employs a covariate, then the underlying response model assumes a covariate-dependent response mechanism (i.e., that the probability of responding is related to the covariate). With such ratio adjustments, the covariate is used to predict the response variable. When the covariate used for adjustment is related both to response propensity

[5] This can be based on known information or the correlation can be calculated.

and the prediction model, both estimation bias and total variance are minimized (Vartivarian and Little, 2002). See Section 4.4 for an example from the Annual Capital Expenditures Survey.

3.3 Additional Methods of Examining Potential Nonresponse Bias

Below are some methods that we have not seen used successfully in practice for business surveys, but could prove helpful in studying the potential for nonresponse bias. See Groves and Brick (2005) for more details.

3.3.1 Compare Nonrespondents and Respondents Across Time

It might be useful to compare early and late respondents on key estimates. In essence this is using late respondents (usually respondents converted after many follow-up attempts) as proxies for nonrespondents. If a mean statistic is significantly different between early and late respondents, this might be an indication of nonresponse bias. (This method has been used for household surveys, but there are questions of whether or not the theory holds.)

3.3.2 Compare Respondent Estimates to Estimates from Administrative Data

Perhaps there are no data available for survey nonrespondents in a given survey, but there is a known total of a key data item from another data source. If we compare this total to the respondents' total, a significant difference might indicate potential nonresponse bias.

3.3.4 Subsample Nonrespondents

In reality, the best way to examine nonresponse bias is to obtain data from a subsample of initial nonrespondents. This would give "real" information about the nonrespondents. Unfortunately, this is too costly (in terms of time, money, and staff) for most programs to consider.

4. Nonresponse Bias Case Studies

This section presents examples illustrating the use of the methods outlined in Section 3 from three case studies: the Monthly Retail Trade Survey (MRTS), the Quarterly Services Survey (QSS), and the Annual Capital Expenditures Survey (ACES). All surveys assume an ignorable response mechanism, where the probabilities of response differ by weighting or imputation cell. MRTS, the QSS, and ACES non-employer companies all assume that the appropriate prediction models and response propensity models were used to minimize bias. These case studies represent a cross-section of our economic programs: MRTS and the QSS use imputation to account for unit nonresponse and have frame data variables highly correlated with their key characteristic, whereas ACES uses weighting adjustments to account for unit nonresponse and has frame data that is inconsistently correlated with its key characteristic.

Note that this section provides a select set of examples from three longer and more complete reports: Rosenthal and Davie (2008), Smith and Thompson (2009), and Lineback (2011).

4.1 Background

4.1.1 Monthly Retail Trade Survey

MRTS is a monthly economic indicator[6] survey. Each month, firms[7] in the MRTS sample are asked to report their sales and inventory data for the month just ending. Estimates of monthly sales and end-of-month inventories are then derived from the collected data.

MRTS is a subset of approximately 12,000 of the 22,000 units in the Annual Retail Trade Survey. The sampling frame for these surveys consists of records extracted for all employer establishments located in the United States and classified under the Retail Trade and Accommodation and Food Services 2002 North American Industry Classification System (NAICS) sectors. The frame is stratified by industry group based on the detail required for publication. Sampling units, in this case Employer Identification Numbers (EINs), are further stratified within industry group by a MOS related to their annual revenue. A new MRTS sample is selected every five years[8].

Sampling units expected to have a large effect on the precision of the estimates are selected with certainty. Within each industry stratum, a cutoff is determined that divides the certainty units from the noncertainty units (Monthly Retail Trade and Food Services Technical Documentation, 2010).

The nonresponse adjustments used by MRTS assume an MAR response model with uniformity within weighting and imputation cells. Imputation cells are defined by NAICS industry cross-classified by MOS quartile. The MRTS MOS variable is highly correlated with monthly sales, but there is not necessarily a strong relationship between the MOS variable and monthly inventories. For more information on MRTS, visit http://www.census.gov/retail/mrts/www/benchmark/2010/pdf/Explanatory_Material.pdf.

4.1.2 Quarterly Services Survey

The QSS provides estimates of total and change in quarterly receipts (published about 75 days after the end of the reference quarter) and early estimates of calendar year receipts for select service sectors. Sampling units for the QSS are groups of businesses under common ownership – generally companies or administratively convenient parts of companies, including EINs. Like MRTS, a new QSS sample is selected every five years[9], and sample units are interviewed each quarter. The QSS sample comprises approximately 6,000 units subsampled from the Services Annual Survey. The MOS variable on the QSS frame is highly correlated with the survey's key item (receipts). For most units on the QSS sampling frame, the MOS is the value of the receipts reported in the most recent Economic Census.

The QSS uses ratio imputation to mitigate unit nonresponse, using the prior period tabulation within imputation cell to predict current period value. Imputation cells are defined by NAICS industry cross-classified by tax-status unless the imputation cell contains fewer than ten respondents. In this case, the imputation cell is collapsed into NAICS subsectors and cross-classified by tax-status. Within each NAICS by tax-status cell, separate imputation cells are created for large companies (mainly consisting of large

[6] Economic indicators are statistical data released to show trends in the economy.
[7] A firm is a group of one or more establishments under common ownership.
[8] The sample is updated quarterly with unit "births" and "deaths".
[9] Like MRTS, the QSS sample is updated quarterly with births and deaths.

businesses selected with certainty) and EINs (primarily consisting of small and medium-sized businesses selected with a design weight greater than one). Imputation parameters are computed from all eligible units in the imputation cell, regardless of certainty status. Further details about the QSS are available at http://www.census.gov/indicator/qss/qsstechdoc.pdf.

4.1.3 Annual Capital Expenditures Survey

The ACES is an annual survey of companies that collects data about the nature and level of capital expenditures by non-farm businesses operating within the United States. Respondents report capital expenditures for the calendar year in all subsidiaries and divisions for all operations within the United States. ACES respondents report total capital expenditures, broken down by type (expenditures on Structures and expenditures on Equipment). The ACES universe contains two sub-populations: employer companies and non-employer companies. Different forms are mailed to sample units depending on whether they are employer (ACE-1) companies or non-employer (ACE-2) companies. Unlike the MRTS and QSS where new samples are selected every five years, new ACE-1 and ACE-2 samples are selected every year so that ACES estimates are based on independent samples. The ACE-1 sample comprises approximately 75% of the total ACES sample.

The ACE-1 frame is developed from administrative payroll data. This auxiliary variable is *not* necessarily highly correlated with capital expenditures. The ACE-1 survey strata are defined by five company size class categories – each based on payroll – within industry: one certainty stratum per industry and four noncertainty strata. The majority of the capital expenditures estimate in a given industry is usually obtained from the certainty and large noncertainty strata; reported zero values for capital expenditures are quite frequent for units from other strata. Thus, the auxiliary data available for the ACE-1 sample are an inconsistent predictor of capital expenditures. There are no corresponding auxiliary/frame variables for the ACE-2 component.

ACES uses "adjustment-to-sample" weighting to account for unit nonresponse (Kalton and Flores-Cervantes, 2003). To do this, sampling weights for unit i (computed as the inverse probability of selection) are multiplied by a weighting-cell specific adjustment factor that is based on data known for both respondents and nonrespondents. For ACES, the weighting cells are the design strata. With ACE-1, the weighting adjustment uses payroll (a frame data variable) as a covariate, i.e., assumes a covariate-dependent response mechanism. The ACE-2 weighting adjustment uses a weighted inverse probability of response (the "quasi-randomization" estimator used with the MAR response mechanism). More details concerning the ACES survey design, methodology, and data limitations are available online at www.census.gov/econ/aces/.

4.2 Response Rate Analyses (MRTS)

Nonresponse bias analysis typically begins by examining response rates over time at the survey level and by select subpopulations. We examined these rates at the program level, the publication level, and by certainty/noncertainty classification. In our case studies, the TQRR analyses are fairly straightforward, since the survey programs publish data for only one or two key characteristics. Statistical analyses of the rates over time account for the stratified sample designs as well as the repeated measures collection, as applicable. The MRTS results presented in this section are representative of all three case studies.

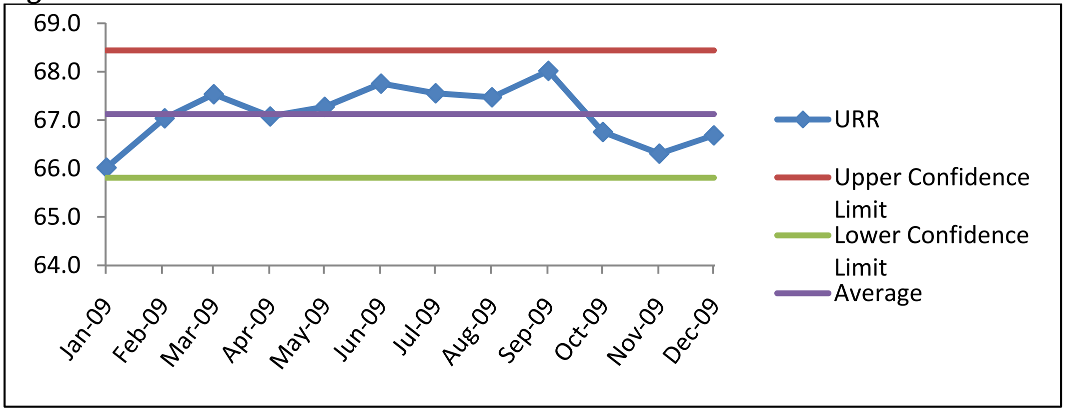
Table 1 presents the URRs for MRTS by certainty status over the studied 12-month period. The program-level MRTS URRs hovered around 67%, with little *visible* deviation (see Figure 2 below). Consistent with the analyst contact procedures described in Section 2.3 above, the URRs for the certainty component are on average much higher than their noncertainty counterparts.

Table 1: URRs for MRTS (1/2009 – 12/2009)

Population	**Avg**	**1/09**	**2/09**	**3/09**	**4/09**	**5/09**	**6/09**	**7/09**	**8/09**	**9/09**	**10/09**	**11/09**	**12/09**
Total	67.1	66.0	67.0	67.6	67.1	67.3	67.7	67.6	67.5	68.0	66.7	66.3	66.7
Certainty	75.9	72.6	74.4	75.1	74.6	76.1	76.4	76.3	77.3	77.6	77.1	77.0	76.5
Noncert	62.7	62.4	63.1	63.5	63.1	62.8	63.4	63.1	62.7	63.4	61.7	61.1	61.9

We examine the URR process using Shewart process control charts. In the Shewart control chart presented in Figure 2, each URR is marked with a diamond. The average rate is indicated by the center asymptote, and upper and lower control limits are obtained by adding and subtracting, respectively, three standard deviations – computed using 12-month averaged rates and sample sizes – to the average rate. From a statistical process control perspective, the URR process is in control with respect to the five criteria provided in Tague (2004), although the mean value of this process falls short of the 80% target.

Figure 2: MRTS Overall URR



Business data program managers are generally more concerned with the TQRRs of their key items than the program-level URR. Again, the justification for this stems from the nature of business survey data, with larger cases contributing substantially to the survey total. From a methodological perspective, this approach is entirely appropriate, as the response mechanism for nonresponding certainty cases is non-ignorable, introducing nonresponse bias. Table 2 presents the TQRRs for MRTS sales by certainty status over the studied 12-month period.

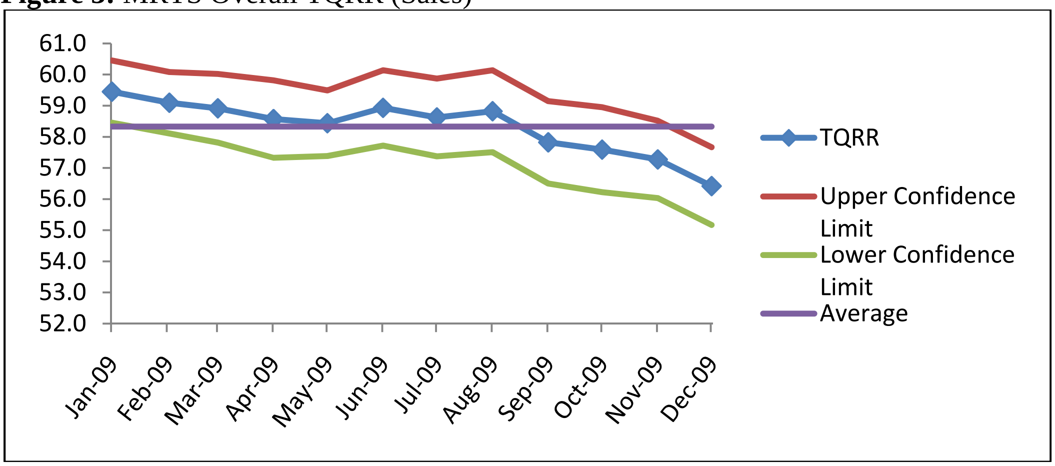
Table 2: TQRRs for MRTS Sales (1/2009 – 12/2009)

Population	**Avg**	**1/09**	**2/09**	**3/09**	**4/09**	**5/09**	**6/09**	**7/09**	**8/09**	**9-09**	**10/09**	**11/09**	**12/09**
Total	78.3	79.1	78.3	78.7	78.5	78.3	78.4	77.7	78.3	77.4	77.9	78.7	79.0
Certainty	94.4	94.5	93.5	94.8	94.9	94.7	95.2	94.2	94.8	93.4	94.2	94.6	94.3
Noncert	58.3	59.4	59.1	58.9	58.6	58.4	58.9	58.6	58.8	57.8	57.6	57.3	56.4

TQRRs for certainty units were consistently high, averaging 94.4%. Further examination of the TQRRs for certainty units by industry sector revealed TQRRs were consistently above 85% in all but one sector. However, the TQRRs in the *noncertainty* components ranged from 30.5% to 62.8%, with an overall program average of 58.3%. Prior to conducting this analysis, the common (mis)perception was that the targeted follow-up of large cases was sufficient to ensure high TQRR. The analyses demonstrated that the noncertainty cases did in fact have a substantial effect on the overall TQRR.

Because the TQRR is a point estimate where the value of the denominator is not constant, it is not appropriate to develop Shewart control charts to monitor this measure. Instead, it is appropriate to use a variation of the *p-chart* – a chart that has been successfully employed by the National Highway Traffic Safety Administration (Pierchala and Surti, 1999). Figure 3 presents such a chart where the control limits (created using random group variance estimates) vary from one sample to the next (i.e., "stairstep" control limits). Since MRTS uses 16 random groups for variance estimation, the critical value is obtained as $t_{15}(90)$.

Figure 3: MRTS Overall TQRR (Sales)



Notice the consistent widths of the confidence intervals over the plotted series. Again, this is not unexpected, given analyst procedures. However, it does provide additional evidence of the stability of the overall data collection procedures for MRTS. Thus, direct interventions (changes) in collection procedures – such as a revised collection instrument (questionnaire) or a revised respondent contact protocol – would be required to change the overall rates.

4.3 Validating the Response Mechanism Using Frame Data (QSS)

After examining response rates by subpopulation, subsequent phases of nonresponse bias analysis investigate the targeted areas "uncovered" by the response rate analyses, delving particularly into response mechanism assumptions.

When highly correlated frame data are available for the variables of interest, we can use these data to examine the assumption of an MAR response mechanism. This section describes our procedures for such investigation, using examples from the QSS analysis. As described in section 4.1.2, the QSS uses a ratio adjustment within imputation cells, where imputation cells are defined by industry (at the six-digit NAICS level) and tax status, and separate imputation cells are created for large companies and EINS. For simplicity, our nonresponse analysis study used certainty status as a proxy for the large company/EIN classification.

To systematically assess the ignorable response mechanism assumption, we performed two-tailed, two-sample *t*-tests of equivalence of the average (mean) frame data variable (census-equivalent receipts) obtained from respondents to the corresponding value obtained from nonrespondents in each imputation cell, restricting our analysis to *only* noncertainty cases. The test statistic within each imputation cell h was computed as

$$t_h^* = \left(\hat{y}_{R,h} - \hat{y}_{NR,h}\right) / \sqrt{\hat{v}\left(\hat{y}_{R,h}\right) + \hat{v}\left(\hat{y}_{NR,h}\right) - 2C\hat{o}v\left(\hat{y}_{R,h}, \hat{y}_{NR,h}\right)}, \quad (3)$$

where
$\hat{y}_{R,h}$ is the Hàjak estimator (Hàjek, 1971) of the respondent imputation cell h mean,
$\hat{y}_{NR,h}$ is the Hàjak estimator of the nonrespondent imputation cell h mean,
$\hat{v}\left(\hat{y}_{R,h}\right)$ is the cell h, random group variance estimates of the respondent statistic,
$\hat{v}\left(\hat{y}_{NR,h}\right)$ is the cell h, random group variance estimates of the nonrespondent statistic, and
$C\hat{o}v\left(\hat{y}_{R,h}, \hat{y}_{NR,h}\right)$ is the cell h, random group covariance estimate.

Since QSS uses 16 random groups, under H_0, $t_h^* \sim t(15)$. Table 3 provides the total number of imputation cells used in each quarter and the total number of cells containing statistically different means for respondents and nonrespondents within the same imputation cell.

Table 3: Comparison of Average Census Equivalent Receipts in QSS Imputation Cells (Quarter 1, 2004-Quarter 4 2005)

Sector		**05Q4**	**05Q3**	**05Q2**	**05Q1**	**04Q4**	**04Q3**	**04Q2**	**04Q1**
Information	Total Cells	12	12	12	12	12	12	12	12
	Significant	1	1	1	2	0	1	2	0
Professional, Scientific, and Technical Services	Total Cells	11	10	11	11	11	11	11	11
	Significant	0	3	2	3	4	3	2	2
Administrative and Support and Waste Management and Remediation Services	Total Cells	5	5	5	5	5	5	5	5
	Significant	0	1	1	1	1	1	0	1
Health Care and Social Assistance (Selected Industries)	Total Cells	4	4	4	4	Not canvassed in 2004			
	Significant	1	2	1	1				

We evaluated the *t*-test results two ways. In each sector, we tested whether the number of cells with significantly different respondent and nonrespondent means was larger than expected due to random variability using binomial tests *when possible* (the binomial test was not feasible in two of the four sectors, each of which contained five or fewer imputation cells). We also looked at the hypothesis test results in the individual imputation cells over time. This examination uncovered *three* of 32 imputation cells with fairly consistent differences between mean respondent and nonrespondent values over the studied eight quarters. These particular subpopulations would be good candidates for additional analyses suggested in Groves and Brick (2005).

The test statistics computed for the QSS were conservative, relying on fairly unstable random group variance estimates. Even so, the results were not unreasonable, with generally slightly more than the expected 10% rejection rate of the null hypothesis that respondent and nonrespondent cell means are the same. Other programs that wish to use these methods could perhaps improve the power of these tests by using averaged estimates of variance and autocovariance in place of individual point estimates. Unfortunately, this was not feasible for the QSS because this program allows imputation cells to vary quarter by quarter due to cell size requirements.

These systematic comparisons rely on parametric assumptions about the imputation cell means. With large samples (even from surveys with complex sample designs) and appropriate variance estimators, these assumptions are likely valid. However, they may not be valid in cases where imputation cells comprise a small number of eligible units (respondent, nonrespondent, or both). The relative variances of the cell means may dwarf the differences in the estimates. This was the case in preliminary frame data analysis for MRTS. As an alternative, nonparametric analysis methods were utilized for their robust properties.

4.4 Validating the Response Models Using Frame and Response Data (ACES)

Weighting adjustments or imputation that use a covariate implicitly assume a prediction model. If the covariate used for adjustment is not related to the survey outcome, then the method used to "correct" nonresponse bias may in fact add bias.

We examined both the response propensity and prediction models for the ACE-1 component of ACES using frame data (payroll) and response data (total capital expenditures), focusing on two issues: 1) the assumption that amount of company payroll is a good predictor of the probability of unit response (propensity model validation); and 2) the assumption that amount of company payroll is a good predictor of reported capital expenditures (prediction model validation).

To examine propensity model assumptions, we fit no-intercept logistic regression models in each noncertainty size strata within industry using the SAS® SURVEYLOGISTIC procedure. For these models, the independent variable is payroll and the dependent variable is the response indicator (I_{hi}). Except for the smallest size-class-strata, these results provide evidence of a relationship between payroll and the probability of responding in the majority of industries/strata.

To examine the prediction model assumptions, we fit no-intercept linear regression models in each noncertainty size strata within industry, again excluding the certainty strata using the SAS® SURVEYREG procedure. For these models, the independent variable is payroll and the dependent variable is total capital expenditures. Again, except for the smallest size-class-strata within industry, these *respondent data based* results provided evidence that the amount of company payroll could be used to predict capital expenditures, reinforcing the validity of the ratio model used for weighting adjustments in three of the four size-class-strata. In fact, for this (large) subset of noncertainty strata, the ACE-1 adjustment is in the desirable situation of using an auxiliary variable that is both related to response propensity and to characteristic outcome.

From these analyses, we concluded that the ratio weighting adjustment methodology used for the three largest size class strata in the ACE-1 design demonstrates the "ideal" properties where the auxiliary variable is related to both the response propensity and the outcome variable (Little and Vartivarian, 2005). We found the opposite in the smallest size class strata in the ACE-1 design. Although we were unable to evaluate whether the respondents comprised a random sample in these strata, in the absence of other related data for modeling, we recommended a change to the "quasi-randomization" weighting procedure used by ACE-2 in these strata.

5. Conclusion

For surveys, it is important to know if and how nonrespondents differ from respondents. This is imperative to knowing if we are making correct inference from sample data.

The OMB directive is to conduct nonresponse bias analysis when a survey program's URR falls below 80%. Even though response rates alone tell little about nonresponse bias, this guideline is a step in the right direction, because it requires us to address the topic and provides a "starting point" for subsequent analyses.

This report serves as a tool for conducting nonresponse bias analysis for business surveys. It addresses the unique nature of business surveys and issues to consider before starting a study. It suggests different approaches based on available data, providing examples from case studies to illustrate the applications.

At the U.S. Census Bureau, we have made substantial inroads in formally establishing methods for conducting nonresponse bias analysis with business survey data. However, our endeavors are by no means complete. We intend to continue exploring new methods for studying and addressing nonresponse bias. In the meantime, we should also be questioning how often nonresponse bias analysis should be conducted, how much bias we are willing to accept in our estimates, and when we should focus on other types of nonsampling error in addition to or in lieu of nonresponse bias. Nonresponse bias analysis, to say the least, is a fruitful area of research for survey methodologists.

Acknowledgements

The authors thank Lance Couzens, Broderick Oliver, Matthew Wheeler, and Benjamin Reist for their help on this project. The authors also thank Rita Petroni and Carol Caldwell for their helpful comments on this paper.

References

Bates, N., Griffin, D., Petroni, R., & Treat, J. (2008). Supporting Document B -Variables, Rates and Formulae for Calculating Response Rates and Reporting Requirements: Economic Surveys and Censuses. Census Bureau guidelines issued 23 December.

Graham, J. (2006). Guidance on Agency Survey and Statistical Information Collections. An Office of Management and Budget memorandum located at http://www.whitehouse.gov/omb/assets/omb/inforeg/pmc_survey_guidance_2006.pdf

Groves, R., & Brick, J. (2005). Practical Tools for Nonresponse Bias Studies. Joint Program in Survey Methodology course notes.

Groves, R., Brick, J.M., Couper, M., Kalsbeek, B., Harris-Kojetin, B., Kreuter, F.,…, & Wagner, J. (2008). Issues Facing the Field: Alternative Practical Measures of Representativeness of Survey Respondent Pools. Found online at http://surveypractice.org/2008/10/30/issues-facing-the-field/.

Hàjek, J. (1971). Comment on "An Essay on the Logical Foundations of Survey Sampling, Part One," In: Godambe, V.P. and Sprott, D.A. (eds.), Foundations of Statistical Inference, Toronto: Holt, Rinehart, and Winston.

Harris-Kojetin, B. (2009). When Should Agencies Conduct Nonresponse Bias Analyses? Presentation at the Federal Committee on Statistical Methodology Workshop, 10 June, Washington, D.C.

Kalton, G., & Flores-Cervantes, I. (2003). Weighting Methods.*Journal of Official Statistics*, **16**, pp. 81-97.

Lineback, J.F. (2011). "Nonresponse Bias Analysis for the Monthly Retail Trade Survey," In progress, internal U.S. Census Bureau memorandum.

Little, R., & Rubin, D. (2002). Statistical Analysis With Missing Data, New York: Wiley.

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

Monthly Retail Trade and Food Services Technical Documentation (2010). Online at http://www.census.gov/retail/mrts/www/benchmark/2010/pdf/Explanatory_Material.pdf.

Ozcoskun, L., & Hayes, M. (2009). Economic Directorate Editing and Imputation Inventory. Internal unpublished U.S. Census Bureau memorandum.

Pierchala, C.E. & Surti, J. (1999). Control Charts as a Tool in Data Quality Improvement. Report no. DOT HS 809 005, National Highway Traffic Safety Administration.

Rosenthal, M., & Davie, J. (2008). Nonresponse Bias Analysis for the Quarterly Services Survey. Internal unpublished U.S. Census Bureau memorandum.

Shao, J., & Thompson, K.J. (2009). Variance Estimation in the Presence of Nonrespondents and Certainty Strata. *Survey Methodology*, **35**, pp. 215-225.

Singh, R., & Petroni, R. (1988). Nonresponse Adjustment Methods for Demographic Surveys at the U.S. Bureau of the Census. Paper presented at the Research Planning Conference on Human Activity Patterns, 10-12 May, Las Vegas, Nevada.

Smith, J.Z., & Thompson, K.J. (2009). Nonresponse Bias Study for the Annual Capital Expenditures Survey. *Proceedings of the Section on Government Statistics*, American Statistical Association.

Tague, N.R. (2004). The Quality Toolbox (2nd Edition), ASQ Quality Press, Retrieved at http://www.asq.org/learn-about-quality/data-collection-analysis-tools/overview/control-chart.html.

Thompson, K.J. (2009). Conducting Nonresponse Bias Analysis for Two Business Surveys at the U.S. Census Bureau: Methods and (Some) Results. *Proceedings of the Section On Survey Research Methods*, American Statistical Association.

Tucker, C., Dixon, J., & Cantor, D. (2007). Measuring the Effects of Unit Nonresponse in Establishment Surveys. Introductory overview lecture at the Third International Conference on Establishment Surveys, American Statistical Association.

Vartivarian, S., & Little, R. (2002). On the Formation of Weighting Adjustment Cells for Unit Nonresponse. Proceedings of the Section on Survey Research Methods, American Statistical Association.