Investigating Nonresponse Subsampling in Establishment Surveys Through Embedded Experiments

Katherine Jenny Thompson[1] and Stephen J. Kaputa

Economic Statistical Methods Division, U.S. Census Bureau
4600 Silver Hill Road, Washington DC 20233

Abstract

Adaptive design strategies for data collection can increase the quality of response data even under a reduced survey budget. The U.S. Census Bureau is investigating nonresponse subsampling strategies for usage in the 2017 Economic Census. Kaputa et al (2014) describes an optimized subsampling procedure for nonrespondent subsampling that selects larger systematic samples in domains that have lower initial response while maintaining approximately equal subsampling intervals, but found that subsampling nonrespondents without changing the data collection procedure may have minimal tangible benefits besides cost reduction. Improving the data collection procedure to target these "hard to reach" subsampled establishments is likely to improve estimates. In this paper, we present the results of a field experiment to test contact strategies for selected small units embedded in the 2014 Annual Survey of Manufactures (ASM). Then we present the design and discuss the proposed analysis strategy for a subsequent embedded experiment in the 2015 ASM, pairing with our proposed nonrespondent subsampling design with the most effective follow-up procedures determined from the earlier test.

1. Introduction

Adaptive design strategies for data collection can increase the quality of response data even under a reduced survey budget. With an adaptive collection design, the data collection procedures can change (*adapt*) during the collection period. Paradata and sample data are used to determine whether and when to change the current procedures (Schouten, Calinescu and Luiten 2013). The overall budget is fixed, but the implementation of a given strategy depends on (1) the realized sample of respondents at a point in time, (2) informative data obtained during data collection about the respondents and nonrespondents, and (3) information known in advance about the survey unit from the sampling frame. Consequently, selecting a probability sample of nonrespondents for follow-up – instead of attempting to contact all nonrespondents – is considered adaptive design, since unit response status (paradata) determines the sampling frame and frame data inform the sample design.

The U.S. Census Bureau is investigating nonrespondent subsampling strategies for the 2017 Economic Census. The specific proposal under consideration is to select a probability subsample of small single unit businesses for nonresponse follow-up (NRFU). Why small businesses? The Economic Census is a quinquennial program whose primary purpose is to provide industry estimates on business and economic items at the national and state levels, as well as in selected metropolitan areas, county, and place levels. The term "census" is a misnomer, as the program includes a probability sample of small businesses in many sectors. However, sampling is somewhat limited due to the geographic publication requirement; with the exception of the construction sector, the lowest sampling rate used is 1-in-20. Even so, business populations are highly skewed, with a few sample units contributing to the majority of the industry totals. Nonresponse follow-up efforts and cognitive research tend to focus on obtaining valid response data from these larger units: see Willimack and Nichols (2010) and Snijkers et al. (2013). Implementing a probability subsample of the larger sample units would doubtless reduce survey costs, as the most expensive NRFU procedures are reserved for these units. However, the reduction in cost would not offset the decrease in estimate quality, as the largest units are often included with certainty and are therefore not well-represented by other units. In

[1] Contact Katherine.J.Thompson@census.gov or Stephen.Kaputa@census.gov. This report is released to inform interested parties of research and to encourage discussion. Any views expressed on statistical issues or operational procedures are those of the authors and not necessarily those of the U.S. Census Bureau.

contrast, smaller businesses tend to contribute much less towards the industry totals, even those cases with large sampling weights.

Given this setting, we consider a *systematic sample* of nonrespondent small businesses that are nonrespondents sorted by a measure of size, a sampling design known to be as efficient as stratified SRS-WOR if the sorted frame is in random order and more efficient if the frame is monotonic increasing or decreasing (Lohr 2010). Kaputa et al (2014) describe an optimized subsampling procedure for nonrespondent subsampling that selects larger systematic samples in domains that have lower initial response while maintaining approximately equal subsampling intervals. This sample design attempts to reduce the effect of nonresponse bias by obtaining a representative respondent sample while simultaneously avoiding overly large increases in variance due to subsampling. Even with an optimized design, subsampling nonrespondents without changing the data collection procedure may have minimal tangible benefits besides cost reduction (Kaputa et al 2014). However, improving the data collection procedure to target these "hard to reach" establishments that were selected for nonresponse subsampling is likely to improve estimates; see Kirgis and Lepkowski (2013).

As mentioned above, survey methods research on data collection strategies and outreach efforts tend to concentrate on the large businesses (companies). In contrast, the small establishments receive very little personal contact (if any) and there is limited cognitive research on preferable contact strategies to draw upon. That said, the literature suggests that there are differences in collected data quality between large and small businesses: see Thompson and Washington (2013), Willimack and Nichols (2010), Bavdaž (2010), Torres van Grinsven, Bolko, and Bavdaž (2014), and Thompson, Oliver, and Beck (2015).

Large companies can operate in many different industries and are often included in more than one business survey. Small businesses often operate in a single industry and are usually selected for one survey by design to reduce response burden. Consequently, cognitive research that relies on focus groups and business site interviews would be useful, but might not necessarily extrapolate to similar sized establishments operating in different sectors. To address this, we conducted an embedded field experiment to test NRFU strategies for selected small units in the 2014 Annual Survey of Manufactures (ASM). This paper presents the findings from this split panel test, focusing on different aspects of response including response propensity, timeliness, and respondent sample balance/quality. Then we present the design and discuss the proposed analysis strategy for an upcoming embedded experiment in the 2015 ASM, which will pair our proposed nonrespondent subsampling design with the most effective follow-up procedures determined from the 2014 test.

2. 2014 Panel Test: Background and Design

2.1. Study Background

The Economic Census is the U.S. Government's official five-year measure of American business and the economy that covers most economic sectors except agriculture. The Economic Census collects a core set of data items from each establishment called general statistics items (examples include annual payroll, total receipts or shipments, and number of employees in the first quarter), as well as information on the revenue obtained from product sales and other industry-specific variables. Although the Economic Census is a single program, the sample design, data collection strategy, imputation models used, and estimation methods can differ by sector. In the upcoming census, data collection will be electronic and use the expanded North American Products Classification System (NAPCS). These anticipated major changes are the catalyst for conducting several embedded experiments on data collection features in several ongoing annual surveys. The results from these tests conducted prior to the 2017 Economic Census will be used to develop the census data collection strategies.

Investigating nonrespondent subsampling and contact strategies for small businesses is part of this overall evaluation process. The presented investigations are embedded into the ASM, an establishment survey designed to produce "sample estimates of statistics for all manufacturing establishments with one or more paid employee(s)." A single-unit (SU) establishment owns or operates a business at a single location, whereas multi-unit (MU) establishments comprise two or more establishments that are owned or operated by the same company. ASM sample units are surveyed for the four years between censuses. Strata are defined by six-digit industry code using the North American Industry Classification System (NAICS). The industry strata are further subdivided into two substrata: certainty and noncertainty. The largest MU and SU establishments in an industry are included with certainty. The remaining establishments are a stratified Pareto-PPS sample, selected with probability proportional to a composite

measure of size and realized sampling weights ranging from 1.01 to 20. The ASM surveys approximately 50,000 establishments selected from a universe of 328,500. About 7,000 of the approximate 20,000 establishments included with certainty are SU establishments, and about 10,000 of the approximate 30,000 noncertainty are SU establishments. To reduce respondent burden, units below a certain threshold do not receive a questionnaire; their data are obtained using administrative records and model imputation. Similarly, the ASM imputes complete records for unit nonrespondents. See http://www.census.gov/manufacturing/asm/ for additional information on the ASM methodology. Note that the ASM survey design is quite different from the majority of the Economic Census sectors, which employ stratified SRS-WOR (not PPS-WOR) designs.

The ASM data collection strategy for SU establishments is very similar to the Economic Census procedures. Furthermore, the ASM questionnaire is a subset of the manufacturing sector's Economic Census questionnaire, the ASM uses the same editing and imputation procedures as the Economic Census, and the ASM collects data from establishments (like the Economic Census). Since we are ultimately concerned with quality effects on collected items from small businesses in the Economic Census, the ASM is therefore an excellent testing ground. Ideally we would want to test Economic Census contact strategies in all economic sectors. Unfortunately, the other annual economic surveys conducted at the U.S. Census Bureau have different sample units (company versus establishment) and collect different items, making the extrapolation to the census a bit less transparent. And, of course not all survey sponsors were comfortable with the risk of affecting mandated reliability levels due to the increased sampling variance caused by subsampling.

For both the ASM and the Economic Census, the collection design varies by type of unit, with contact strategies designed to ensure that the largest cases provide valid response data. Nonresponding MU establishments and the larger SU establishments[2] receive more frequent NRFU contacts and can include more expensive personal contacts, such as phone follow-up. The remaining small single-unit establishments receive reminders, but are very unlikely to receive personal contact. The contact strategy for single units in the 2014 ASM relied entirely on mail outreach. Altogether, there are five possible contacts:

- Initial contact letter (all sampled units), providing a deadline and requesting internet response via a secure system that includes username and password
- 1st NRFU: reminder letter stating that response is past due and again requesting response via internet.
- 2nd NRFU: reminder letter plus paper questionnaire. Historically, the questionnaire was included in the initial mailing; this is the first time the questionnaire mailing has been delayed until the second NRFU. This was done primarily to push internet collection over paper collection, but also to save money.
- 3rd and 4th NRFU: progressively threatening letters requesting response. See the Appendix for a sample.

Historically, the Economic Census and the ASM have been mail-out/mail-back collections. However, the U.S. Census Bureau is strategically increasing the use of internet (web) data collection in both household and business surveys over other modes of data collection. The 2017 Economic Census will be an all-internet collection. Web data collection affords substantive cost reductions and is believed to improve data quality over paper collections (Thompson, Oliver, and Beck 2015). Although the majority of the surveys conducted in the economic directorate of the Census Bureau are moving towards complete internet collection, there is some concern that not all businesses may be able to respond through this mode, along with a concern that response may be affected by previous response conditioning towards paper collection. Especially because of the latter, the ASM subject matter experts were not comfortable with a complete push towards web collection for the 2014 survey year. Instead, they preferred to "transition" respondents by including a paper questionnaire at some stage of the NRFU. From the 2015 collection onward, the ASM NRFU procedures will abandon this procedure.

Previous research conducted in the 2007 and 2012 Economic Census demonstrated that response rates for SU establishments in low responding areas are improved by including a certified mailing in the NRFU protocol (Marquette, Kornbau, and Toribio 2015). In the cited study, the certified mail reminder was reserved for selected industries with low response remaining after the 2nd or 3rd NRFU attempt. Our experiment expands the target population to all single unit establishments, not just those located in low responding industries.

[2] Included with certainty in the Economic Census or the ASM.

Supplemental material that highlights due dates, mandatory response and survey utility can also affect response. These messages can be framed positively or negatively. Positive material tries to motivate response through a promotional message illustrating the importance of the collected data to national estimates/decisions or to the designated establishment e.g., a flyer listing the uses of the survey's data in public policy decisions or providing interesting facts about the industry. Alternatively, negative material carries a threatening message, such as legal action for mandatory surveys. Although this list of strategies is certainly not exhaustive, they are common contact strategies with limited budget effects. For example, personal phone calls and more frequent contacts have been shown to be effective for eliciting response (Marquette, Kornbau, and Toribio 2015), but are quite expensive and were not considered in this study.

2.2. Experimental Design

The 2014 test includes *all* SU establishments, regardless of certainty/noncertainty status. Three separate NRFU strategies were tested. To obtain the treatment panels, the ASM SU sample was blocked on three-digit NAICS industry and certainty/noncertainty status. Establishments were sorted within blocks by the frame measure of size (MOS) and systematically assigned to treatment panels using a random start. Thus, each treatment panel's composition is balanced by industry, certainty/noncertainty status, and establishment size. At the time of initial mail, each panel contained approximately 5,700 single units and was expected to have 3,300 nonrespondents at the time of the second NRFU.

Table 1 presents the tested contact strategies by treatment panel. Each round of contact followed the same fixed calendar schedule. The experimental treatments (NRFU strategies) that do not follow the normal contact strategy appear in red. For all panels, the initial contact and 1st NRFU procedure are the same. The control panel uses the contact strategies for the 2014 ASM described in Section 2.1. Treatment panels 1 and 2 respectively test the effectiveness of a certified letter (T1) and the negative flyer/letter provided in the Appendix (T2) as the 2nd NRFU. All NRFU protocols include a questionnaire mailing (denoted as form in Table 1) for the reasons outlined in Section 2.1. However, this mailing is delayed until the 3rd NRFU contact for the T1 and T2 panels. Lastly, all treatment panels received the same final letter as 4th NRFU contact.

Table 1: Contact Strategies by Treatment Panel

Panel	**Initial mail**	**1st NRFU**	**2nd NRFU**	**3rd NRFU**	**4th NRFU**
Control (C)	Letter	Letter	Form	Letter	Letter
Treatment 1 (T1)	Letter	Letter	**Certified Letter**	**Form**	Letter
Treatment 2 (T2)	Letter	Letter	**Letter/ Negative Flyer**	**Form**	Letter

3. Results

The ultimate goal of pairing a probability subsample of nonrespondents with a targeted NRFU strategy is to reduce the nonresponse bias in the survey estimates. Each panel of the 2014 ASM test is a random subsample of the full ASM sample of SU establishments. It is possible that a given NRFU strategy could improve response rates without comparable improvements in estimate quality if all three sets of respondent samples are "representative" (i.e. all sample units' response mechanism is missing completely at random or missing at random) or are all equally lacking in a subdomain (e.g., the smallest establishments). On the other hand, it is not difficult to determine a single "treatment effect" when one NRFU method exhibits improved performance on one measure of nonresponse bias over the others.

Andridge and Little (2011) observe that there are three components that can be used to assess the potential for nonresponse bias: the amount of nonresponse, the differences between respondents and nonrespondents on fully observed characteristics (e.g., paradata, frame data), and the relationship between these fully observed characteristics and the survey outcomes (only measureable among respondents). Wagner (2012) presents a useful typology for alternative indicators for the risk of nonresponse bias that incorporates this framework: (1) indicators involving the response indicator; (2) indicators involving the response indicator and frame data or paradata; and (3) indicators involving the response indicator, frame data or paradata, and the survey data. The analyses below are categorized with this typology.

3.1. Type (1) and Type (2) Indicator Analyses
3.1.1. The Effects of Different NRFU Treatments on Response

For business surveys, unit response rates are computed as unweighted ratios of respondents to eligible cases. This avoids overrepresentation of the smaller cases with larger weights in the response rate. For computation of the official rates, a respondent is defined as eligible reporting unit for which: (1) an attempt was made to collect data; (2) the unit belongs to the target population; (3) and the unit provided sufficient data to be classified as a response (Thompson and Oliver 2012). In our case study, the data have undergone minimal editing. Consequently, we examine a "proxy" response rate, categorizing a unit as responding if it provided a value for annual payroll; this value may or may not be used in the final tabulations. For simplicity, hereafter we refer to this proxy rate as the unit response rate. Figure 1 plots unit response rates over time by treatment panels for certainty cases within panel,= and for noncertainty cases within panel.

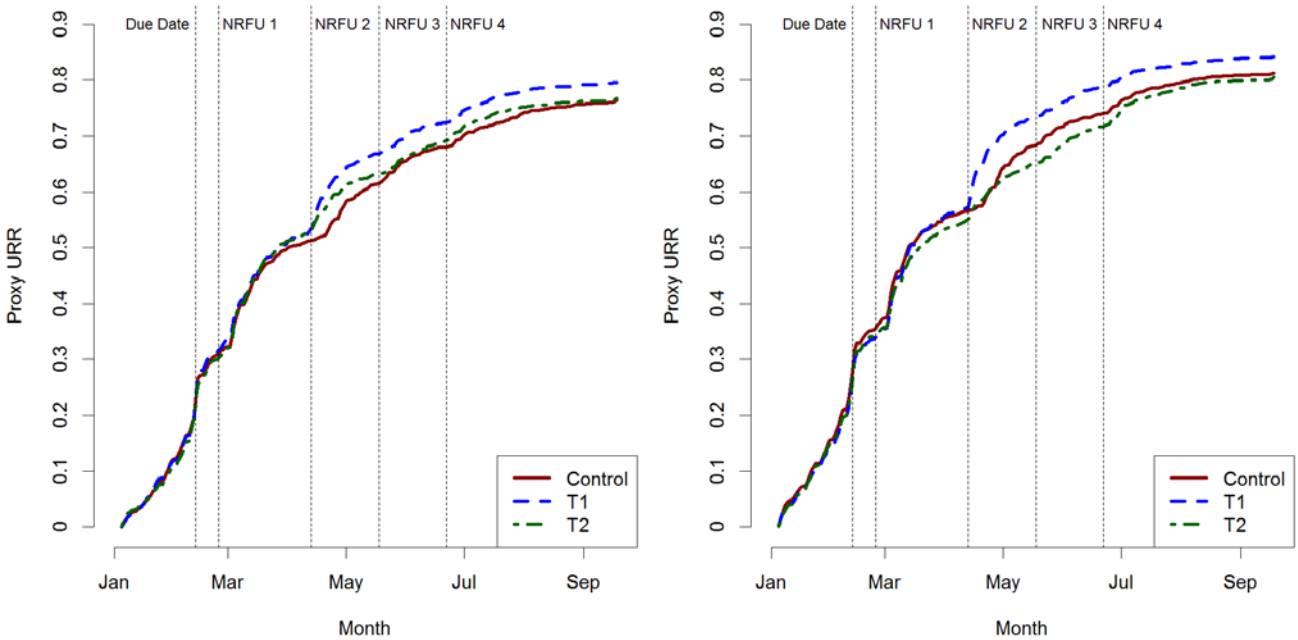

Figure 1: Proxy Response Rates by Treatment Panel (Certainty and Noncertainty)

Regardless of certainty status, these plots provide strong visual evidence of an improved response rate when the certified letter is used for the 2nd NRFU attempt, in contrast to the other treatments, whose response rate plots are indistinguishable.

We test the hypothesis of independence of treatment and <u>overall</u> response rate using chi-squared tests for independence for the complete set of test cases and by certainty and noncertainty status. Although the official unit response rates are computed without sampling weights, testing for differences without incorporating complex design features of the stratified PPS-WOR sample can lead to erroneous conclusions (Rao and Scott 1987). Using the Rao-Scott adjusted test implemented in PROC SURVEYFREQ addresses this problem (SAS/STAT(R) 9.3 User's Guide 2015), but incorporates the sampling weights, yielding slightly different distributions than the official ones. The SAS procedure incorporates the fpc-adjustment needed in the noncertainty subdomains to account for the reduction in sampling variance from without-replacement sample. Consequently, we conducted both unweighted and weighted analyses for the noncertainty case evaluations and unweighted analyses only for the certainty case evaluations. The unweighted analyses use the standard Pearson statistics; the weighted analysis use the Rao-Scott adjusted test, which incorporates the complex survey design features. First, we tested for equivalence of final unit response rate by treatment panel. Table 2 presents these results by subdomain.

Table 2: Tests for Equivalence of Unit Response Rate by Treatment

Treatment Panel		Certainty	Noncertainty	
		Unweighted	Unweighted*	Weighted**
C	Response Rates	69.80	78.90	76.97
T1		73.21	81.73	80.86
T2		69.57	78.10	75.85
	Test Statistic	9.50	15.05	11.48
	p-value	0.0087	0.0005	0.003

*$\chi^2(2)$, Pearson test
** $\chi^2_{RS}(2)$, Rao-Scott Adjusted test

All test reject the null hypothesis (at α=0.05), providing evidences that at least one treatment results in a different unweighted response rate, with an "across the board" effect (i.e. not confined to a single subdomain).

However, the first round of NRFU is the same for all treatments, and it is possible that the differences in unit response rates could be attributable to a pre-existing difference in unit response rates between panels that existed before the 2nd round of follow-up. To assess this, we performed similar analyses within subdomain, treating the cases that responded <u>before</u> the second round of NRFU as the respondents. None of these tests provide sufficient evidence (at α=0.05) of a "treatment panel effect" on the unit response rate before the 2nd follow-up procedure is introduced. It does appear that the difference in response rates between treatment panels is primarily attributable to the differences in NRFU procedures at the 2nd round of NRFU.

Inspecting Table 2, it appears that the unit response rates for the control panel (C) and the letter/flyer (T2) panels are very similar. Performing reduced tests – dropping the T1 treatment panel – yields no treatment effect for all SU cases, providing further evidence of no treatment effect using the negative letter (unweighted $\chi^2(1) = 0.48$, p-value = 0.49; weighted $\chi^2_{RS}(1) = 0.44$, p-value = 0.51), for certainty cases (unweighted $\chi^2(1) = 0.02$, p-value = 0.87), and for noncertainty cases (unweighted $\chi^2(1) = 0.62$, p-value = 0.42; weighted $\chi^2_{RS}(1) = 0.49$, p-value = 0.49). Ultimately, these results provide evidence that sending a certified letter, followed by remailing the form to the remaining nonrespondents, increases the response rate over the other two treatments. Equally important, these results provide evidence that sending a strongly worded message of delinquency and obligation, followed by remailing the form to the remaining nonrespondents, does not improve the response rate of these "low-probability respondent" cases.

3.1.2. The Effects of Different NRFU Treatments on Length of Time to Respond

Time to respond can be modeled as failure-time data, where failure-time is defined as length of time to respond. Using survival analyses methods, we fit Cox proportional hazards regression models to the failure-time data, assuming that the dependent variable (time to respond) is continuous. Our objective is to model the hazard function[3] ($\lambda_i(t)$) for each unit i at time t as $\lambda_i(t) = \lambda(t; Z_{i(t)}) = \lambda_0(t)\exp(Z_i'(t)\beta)$, where $\mathbf{Z}_i(t)$ is a vector of explanatory variables at time t for unit i and $\boldsymbol{\beta}$ is the associated unknown vector of regression parameters, assumed to be the same for all individuals (Cox 1975). The primary statistic of interest is the hazard ratio: a value larger than one indicates a positive effect on response due to a treatment, whereas a value less than one indicates a negative effect.

To account for the complex survey design, we use PROC SURVEYPHREG to predict the onset of response and regress on treatments (SAS/STAT(R) 9.3 User's Guide 2015) with the control panel treated as a baseline. Of course, as with the previous analysis, the fpc-adjustment is needed for the noncertainty subdomain, but cannot be used for the certainty subdomain (we use PROC PHREG). We examine failure-time as a function of treatment panel (baseline=control). Table 3 provides the maximum likelihood estimates for each parameter along with the associated hazard ratio. Statistically significant regression parameters and hazard ratios at α=0.05 are indicated by an asterisk.

[3] the instantaneous rate at which a unit will respond, given that the unit has not already responded

Table 3: Proportional Hazards Regression Model Tests by Subdomain

	Certainty		Noncertainty	
Parameter	Estimate	Hazard Ratio	Estimate	Hazard Ratio
T1 (Certified Letter – Form)	0.06	1.07	0.09*	1.10*
T2 (Letter/Flyer – Form)	0.08*	1.08*	-0.03	0.97
Contrast (T1-T2) test p-value	-0.36		<0.0001	
Global Test p-value	0.0542		<0.0001	

For the noncertainty SU cases, these results provide evidence of a positive (and significant) effect on time to respond for the cases in the certified letter treatment (T1) panel and an increased probability of responding over the current procedure. This increased probability of responding has an effect on the noncertainty SU cases' overall response rates, as seen in the previous section. These units are responding more quickly and at a higher rate. These results provide evidence that the certified letter treatment is more effective than the current procedure for eliciting responses from the noncertainty SU cases, with further confirmation provided by the pairwise contrast test of the T1 and T2 treatments in the noncertainty population (p-value = 0.0001).

For the certainty SU cases, there is likewise evidence of a positive a positive (and significant) effect on time to respond for the cases in the certified letter treatment (T2) panel and an increased probability of responding over the current procedure. Recall, however, that there is no significant different in unit response rate for these units in the control and T2 panels. Thus, the increased probability of eventually responding has no practical effects, and it appears that the responding units are simply answering the survey in a timelier manner.

3.1.3. The Effects of Different NRFU Treatments on Representativeness of the Respondent Sample

The previous analyses provided evidence that the T1 NRFU strategy elicits more responses from small (noncertainty units) than the other two NRFU strategies under consideration. If this is the case – and the other two panels *underrepresent* these smaller units in the response set – then the T1 sample composition should be a more "representative" subsample of the ASM SU sample. We explore this hypothesis using two indicators of representativeness defined in Särndal and Lundquist (2014): the balance indicator and the distance indicator. 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.

Of course, obtaining a representative respondent sample would be a goal of nonrespondent subsampling (not considered here). For this analysis, it is simply important to see if one particular treatment yields a more representative sample than the others or if all have equally balanced or equally unbalanced samples.

Let
y = characteristic of interest, subject to nonresponse
x = auxiliary variable available for all sampled units
P = weighted response rate = $\sum_{i\in s} w_i I_i / \sum_{i\in s} w_i$, where I_i is a unit response indicator and w_i is the design weight

Assume that $y \approx \beta x + \varepsilon$. Following Särndal and Lundquist (2014), define

Balance $B_x = \bar{x}_r - \bar{x}_s$, the difference between mean value for respondents and mean value for sampled units where $\bar{x}_r = \sum_{i\in s} w_i x_i I_i / \sum_{i\in s} w_i I_i$ and $\bar{x}_s = \sum_{i\in s} w_i x_i / \sum_{i\in s} w_i$. $B_x = 0$ is an indicator that the respondent set is a random sample of the parent sample for all collected variables correlated with x, the auxiliary variable.

Imbalance Measured as $IB_x = (\bar{x}_r - \bar{x}_s)' \Sigma_s^{-1} (\bar{x}_r - \bar{x}_s)$, $\Sigma_s = \sum_{i\in s} w_i x_i x_i' / \sum_{i\in s} w_i$.

A *balance indicator for variable x* is given as $BI_x = 1 - 2P\sqrt{IB_x}$. This measure is bounded between 0 and 1, with values close to 1 indicating balance on the respondent sample for the studied variable. However, it tends to overestimate this quality.

Distance $D_x = \bar{x}_r - \bar{x}_{nr}$, the difference between mean value for respondents and mean value for nonrespondents on variable x. This is measured as $D_x = (\bar{x}_r - \bar{x}_{nr})'\Sigma_s^{-1}(\bar{x}_r - \bar{x}_{nr})$, and is bounded by $1/\sqrt{P(1-P)}$.

Ideally, the balance indicator should be near 1 and distance indicator should be near 0. Table 4 presents the distance and balance measures on 2014 administrative payroll and unit measure of size from the sampling frame (MOS) by panel and certainty status (subdomain).

Table 4: Distance and Balance Measures by Panel on Administrative Payroll and Measure of Size

	Certainty				Noncertainty			
	Payroll		MOS		Payroll		MOS	
Treatment	BI_x	D_x	BI_x	D_x	BI_x	D_x	BI_x	D_x
C (Control)	0.98	0.00	1.00	0.00	0.97	0.01	0.98	0.00
T1 (Certified Letter – Form)	0.98	0.00	0.97	0.01	1.00	0.00	0.99	0.00
T2 (Letter/Flyer – Form)	1.00	0.00	1.00	0.00	0.99	0.00	1.00	0.00

Overall, all panels are in balance with respect to both auxiliary variables and the corresponding distance measures are very close to zero, regardless of certainty status. For the certainty units, the T2 panel (Letter/Flyer – Form) is perfectly balanced on both auxiliary variables and has the minimal difference (at the fourth decimal place, not shown). For the larger single unit establishments, this appears to support the subject matter experts' contention that the harsh tone of the flyer elicits response at higher rate than the other NRFU protocols. This is not the case for the noncertainty units.. Here, the two alternative treatments (T1 and T2) result in more balanced respondent samples than that obtained with the current (C) procedure. One could argue that the marginally improved results on 2014 administrative payroll with the T1 NRFU procedure are more relevant than the reverse seen with the T2 NRFU procedure on the T2 sample because the administrative data are obtained from the concurrent collection period. Again this is not a strong argument – or completely convincing evidence of superior balance for either treatment – given the optimality of the balance and distance indicators for all treatments and panels.

3.2. Type (3) Indicators

Recall that Type 3 indicators combine response rates, frame or auxiliary data available for all sampled units, and survey data to model potential effects of nonresponse bias. The proxy pattern-mixture (PPM) analysis approach first proposed by Andridge and Little (2011) falls into this indicator category. In brief, the PPM model reduces a set of fully observed auxiliary variables to a single "proxy" variable X. The joint distribution of a survey outcome Y and this proxy X is modeled as a bivariate normal distribution with separate parameters for respondents and nonrespondents (a pattern-mixture model). Andridge and Thompson (2015B) develop a PPM model using a bivariate gamma model that is more appropriate for the studied skewed business populations. Either formulation produces adjusted estimates of the mean of Y under different missingness mechanisms, explicitly specified in the model used to link the proxy and outcome variable.

The fraction of missing information (FMI) has been proposed as a metric for assessing the risk of nonresponse bias for a specific adjusted survey estimate (Wagner 2010, Wagner 2012, Andridge and Little 2011, Andridge and Thompson 2015 (A and B)). The FMI is a measure of loss of precision due to nonresponse, and is the ratio of between-imputation variance to total variance for a specific estimator (Little and Rubin 2002). The FMI value for a given Y is bounded between 0 and 1, with a value close to zero indicating little or no nonresponse bias effects in the variable after adjustment and a value close to one indicating the reverse. In the PPM framework, FMI is computed with respect to an assumed response mechanism. To assess the sensitivity of the computations to this, we compute the FMI at the two extremes, specifically missing at random (MAR) and not missing at random (NMAR). If the FMI values for the variable obtained under different response mechanisms are close together, then the inflation of variance due to an MNAR mechanism is not severe, relative to the MAR mechanism. For a more detailed discussion of the factors impacting FMI and its use in the PPM framework, see Andridge and Thompson (2015A).

Using the bivariate-gamma PPM formulation presented in Andridge and Thompson (2015B), we compare the FMI within treatment panel on three separate survey items (payroll, total employment, and receipts), producing a separate proxy for each by regressing the outcome variable on frame MOS within 3-digit industry (a no-intercept linear regression model). As recommended by Andridge and Thompson (2015A and B), we use multiple imputation to produce all estimates, with 200 draws given a burn-in period of 500 draw and thinning at every 10th draw.

Wagner (2010) computes the FMI of several key variables during survey collection to study whether additional data collection decreases nonresponse bias effects over time. Our evaluation is analogous, and we are particularly interested in seeing whether the increased response rates in the T1 panel have a corresponding beneficial effect in terms of nonresponse bias reduction on more than one variable. Furthermore, examining the FMI for an item by treatment panel while holding the prediction model constant provides insight into the respondent data population, building on the balance and distance indicators presented in Section 3.1.3 by examining collected survey data. If the respondent set is "balanced," we expect to see low FMI and we hope to see small differences in corresponding FMI estimates (MAR vs. NMAR). Of course, the FMI values are strongly related to the strength of the predictors used in the proxy. In some instances, a treatment effect might be completely ameliorated by an excellent predictor (strong proxy). The converse can also be true if the relationship between predictors and outcome is not strong (weak proxy). Consequently, we examine three items with varying proxy strength: the payroll proxy fit is extremely strong (as expected), the employment proxy is very strong, and the receipts proxy fit is weak.

Figure 2 presents the adjusted-R^2, nonresponse rate, and FMI values by treatment panel for the three studied items. Here, we focus exclusively on SU noncertainty cases, as this is the only subpopulation eligible for study in the 2015 ASM test. The models used to develop the payroll and employment proxies use all respondent data; five extremely large outliers were removed from the total receipts model to improve the fit [Note: these outliers would not be present in fully edited data].

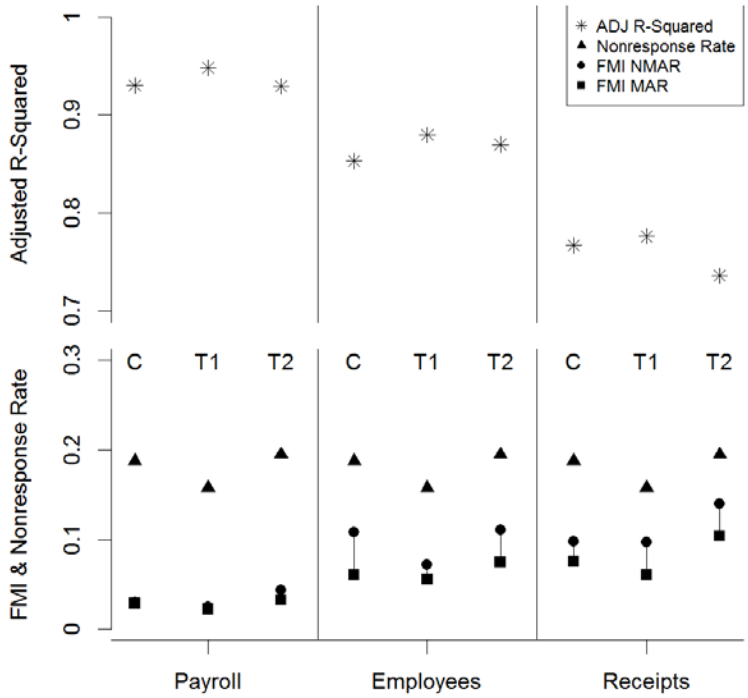

Figure 2: FMI for Noncertainty Single Units

Regardless of outcome variable, Figure 2 shows that all FMI values for all variables are well below the nonresponse rate, signaling that the studied estimates are not overly subject to nonresponse bias in any given panel. However, the T1 panel estimates always have the lowest FMI of the three corresponding panel estimates under the MAR and NMAR response mechanisms. Since the regression models differ only in the treatment variable, this provides

limited evidence that the T1 respondent data set is more representative on the studied variables than the others given the assumed models. Furthermore, with the exception of receipts, the T1 FMIs exhibit the smallest spread. With payroll, this is likely a consequence of the extremely strong proxy. However, with employees, the reduced spread shows the combined effect of the fairly strong proxy and the small distance between respondents and nonrespondents. With receipts, the T1 FMI levels remain smaller than the T2 levels; the similar spread shows the effect of the weaker proxy on correcting nonresponse.

Not surprisingly, the balance and distance indicator analyses provided in Section 3.1.3 are nearly identical to the FMI results for the payroll model given that the ASM measure of size variable is largely derived from Economic Census payroll values. Restricting this analysis to these balance and distance indicators – or to the FMI from the payroll proxy – appears to be overly optimistic about the effect of NRFU protocol on respondent sample composition. Examining the FMIs of the employment and receipts proxies is more revealing, as they demonstrate a negative effect on the respondent sample for other – more difficult to impute – items obtained using T2 NRFU protocol.

4. Proposed 2015 ASM Test

The primary purpose of the 2014 embedded experiment was to determine a NRFU protocol that elicited improved response from small businesses sampled in the ASM. The larger research question is how to effectively implement this protocol in an adaptive collection design, balancing competing interests of costs and quality. One considered approach is the selection of a probability subsample of small single unit businesses for NRFU; an alternative approach is to continue the current procedure of NRFU of all originally sampled units, but target the more expensive procedures to subdomains that have lower initial response. The 2015 ASM tests will consider both approaches, serving two purposes:

1. Compare quality effects of using targeted selection of nonrespondents to receive certified mail reminder compared to sending all nonrespondents a certified mail reminder letter (adaptive design NRFU protocol versus fixed design NRFU protocol)
2. Compare quality effects of selecting a probability subsample of nonrespondents for NRFU

As with the 2014 test, the target population will be the ASM SU nonrespondents. The experimental design is different, however. All ASM noncertainty SU cases receive the same initial contact letter, due date reminder letter, and 1st NRFU letter. This maximizes the usage of previously-proven contact strategies for this survey. After the 1st NRFU concludes, the ASM industries will be split into two separate panels, based on blocking criteria such as sample size, percentage of noncertainty sample, and historic or proxy unit response rates [Note: the final blocking criteria is not available before the survey is conducted, although it is possible that historical rates could be used preliminarily]. All nonrespondent units in the Control panel will receive a certified letter reminder (2nd NRFU) and an Office of General Council (OGC) letter if they have not responded to the 2nd NRFU attempt.

The treatment panel assignment is more complex. We will use the optimized allocation described in Kaputa et al. (2014) to select a targeted systematic probability subsample of units in each industry (T1). This allocation strategy attempts to equalize subsampling rates while maintaining target response rate levels in each industry; the end result is that sampling rates are higher in industries with low response rates. See Appendix Two for details on this allocation procedure. This will be a *systematic sample* of nonrespondent small businesses sorted by MOS, a sampling design known to be as efficient as stratified SRS-WOR if the sorted frame is in random order and more efficient if the frame is monotonic increasing or decreasing (Lohr 2010).These sampled units (T1) will receive a certified letter reminder (2nd NRFU) and an OGC letter if necessary. The remaining complimentary units will receive a reminder letter (not certified) and an OGC letter if necessary. Consequently, *all* nonresponding units in the treatment panel will receive *some* form of NRFU. See Figure 3 for an illustration.

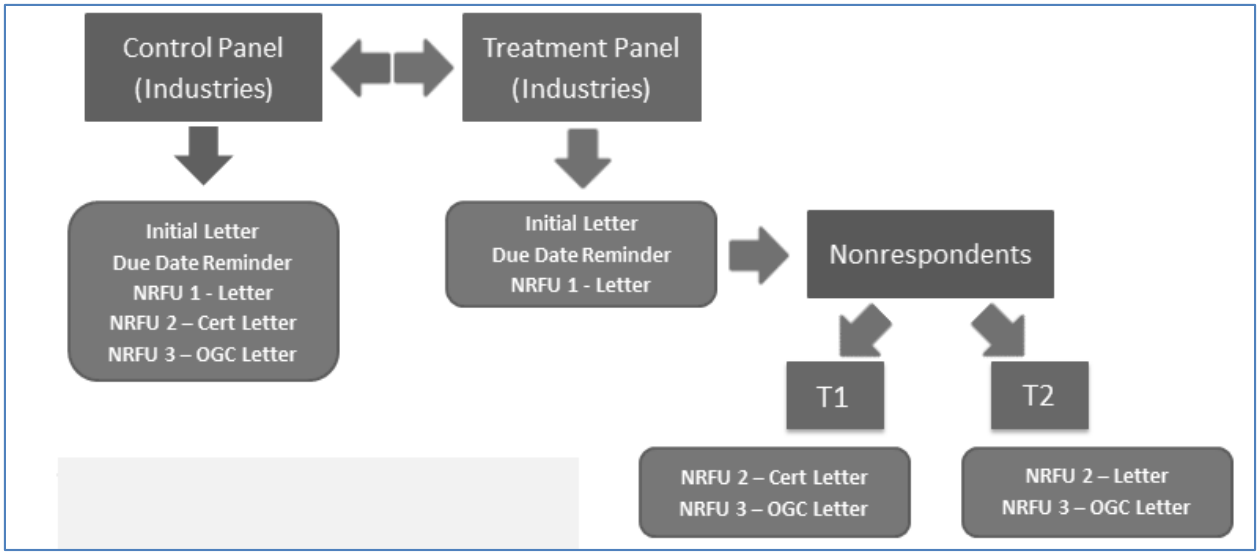

Figure 3: Experimental Design for 2015 ASM Adaptive Design Test

By retaining the subsampling parameters associated with the T1 selection, comparisons can be made between the Control panel and the T1 (subsampled units) panel, as well as the Control panel and the Test (T1 + T2) panel. The surveying costs with the Test panel will be slightly less than 50% of the Control panel costs, as the allocation will use an overall 1-in-2 sampling rate. Retaining the subsampling weights will allow us to simulate the effects on survey estimates of using only a probability sample of nonrespondents. By surveying all units, we can examine the effects on quality of using a less expensive but adaptive data collection protocol.

5. Conclusion

As methodologists, we often rely on simulation studies to assess competing research proposals. This can work well when the underlying conditions are constant, e.g. comparing alternative sampling designs on the same frame. If the underlying conditions can change – or are random variables themselves – then using simulation results as final decision rules can be risky.

Our original research goal was to determine a method of subsampling nonrespondents from a hard-to-reach population with known low response propensities. Treating this subject as an adaptive survey design problem, we developed an allocation strategy that "limited the damage" caused by variance inflation due to subsampling. Even so, we advocated researching alternative estimators and alternative contact strategies. In the latter case, simulation simply was neither viable nor useful. Real experimentation – on the actual target audience – was required. Complicating the experiment was the dearth of available research results on NRFU strategies for small businesses. Subject matter experts had anecdotal opinions (all varying), but no quantitative results.

The presented case study proves the value of an embedded experiment in this situation. The test was not difficult to conduct and provided convincing evidence to the subject matter experts. The split panel design allowed for analysis on a variety of diverse small businesses in different industries, as opposed to focus groups. Our statistical analyses were compromised somewhat by the challenges of small sample sizes, complex survey design effects, and ongoing production. We did abandon many of our original analyses whose power was detrimentally affected by large sampling variances. Nevertheless, the remaining analyses presented here provided convincing results that we believe can be extrapolated to other similar populations. Best of all, these analyses go beyond simple response rate comparisons and this framework can be applied in general to other embedded experiments.

Finally, the case study presented in this paper was a dress rehearsal for a more complex test. The production experts have begun modifying their systems for other tests, and the program coordinators have metrics for ensuring that planned design is used. With these considerations fresh, planning is well underway for the 2015 ASM test.

Acknowledgments

The authors thank Cha-chi Fan, Carma Hogue, and Eddie Salyers for their useful comments on earlier versions of this manuscript, Laura Bechtel for her invaluable contribution to the allocation research, and Robert Struble for designing and implementing the panel selection procedures for the 2014 and 2015 ASM tests.

References

Andridge, R.R. and Little, R.J.A. 2011."Proxy Pattern-Mixture Analysis for Survey Nonresponse." *Journal of Official Statistics* 27: 153-180. Available at: http://www.jos.nu/Articles/abstract.asp?article=272153.

Andridge, R.R. and Thompson, K.J. 2015(A). "Using the Fraction of Missing Information to Identify Auxiliary Variables for Imputation Procedures via Proxy Pattern-Mixture Models." *International Statistical Review* 83(3): 472–492. DOI: 10.1111/insr.12091.

Andridge, R.R. and Thompson, K.J. 2015(B), accepted. "Assessing Nonresponse Bias in a Business Survey: Proxy Pattern-Mixture Analysis for Skewed Data." *Annals of Applied Statistics.*

Bavdaž, M. 2010. "The Multidimensional Integral Business Survey Response Model." *Survey Methodology* 36: 81- 93.

Berthelot, J.M. and Latouche, M. 1993. "Improving the Efficiency of Data Collection: A Generic Respondent Follow-up Strategy for Economic Surveys." *Journal of Business and Economic Statistics* 11(4): 417-424

Cox, D. R. 1975. "Partial Likelihood." *Biometrika* 62: 269–276.

Kaputa, S.J., Bechtel, L., Thompson, K.J., and Whitehead, D. "Strategies for Subsampling Nonrespondents for Economic Programs." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 6, 2014. Available at: http://www.amstat.org/sections/srms/Proceedings/311881_88310.pdf.

Kirgis, N. and Lepkowski, J. 2013. "Design and Management Strategies for Paradata-Driven Responsive Design: Illustrations for the 2006-2010 National Survey of Family Growth." In *Improving Surveys With Paradata*, edited by Frauke Kreuter. New Jersey: John Wiley & Sons, Inc.

Little, R.J.A. and Rubin, D.B. 2002. *Statistical Analysis with Missing Data (2nd Edition)*. New York: Wiley.

Lohr. S.L. 2010. *Sampling: Design and Analysis (2nd Edition).* Boston: Brooks/Cole.

Marquette, E., Kornbau, M., and Toribio, J. "Testing Contact Strategies to Improve Response in the 2012 Economic Census." In Proceedings of the Section on Government Statistics: American Statistical Association, August 10, 2015.

Rao, J.N.K. and Scott, A.J. 1987. "On Simple Adjustments to Chi-Square Tests with Sample Survey Data." *The Annals of Statistics* 15(1): 385-397.

Särndal, C. and Lundquist, P. 2014. "Accuracy in Estimation with Nonresponse: A Function of Degree of Imbalance and Degree of Explanation." *Journal of Survey Statistics and Methodology* 2(4): 361-3087.

"SAS/STAT(R) 9.3 User's Guide." *SAS/STAT(R) 9.3 User's Guide*. N.p., n.d. Web. 09 Oct. 2015.

Schouten, B., Calinescu, M., and Luiten, A. 2013. "Optimizing Quality of Response through Adaptive Survey Designs." *Survey Methodology* 39(2): 29-58.

Snijkers, G. Haraldsen, G., Jones, J.,and Willimack, D. K., editors. 2013. *Designing and Conducting Business Surveys*. New Jersey: John Wiley & Sons, Inc.

Thompson, K.J. and Oliver, B.E. 2012. "Response Rates in Business Surveys: Going Beyond the Usual Performance Measure." *Journal of Official Statistics* 28: 221-237. Available at: http://www.jos.nu/Articles/abstract.asp?article=282221.

Thompson, K.J., Oliver, B., and Beck, J. 2015. "An Analysis of the Mixed Collection Modes for Two Business Surveys Conducted by the US Census Bureau." *Public Opinion Quarterly* 79 (3): 769-789. DOI: 10.1093/poq/nfv013.

Thompson K. J., and Washington K. T. 2013. "Challenges in the Treatment of Unit Nonresponse for Selected Business Surveys: A Case Study." *Survey Methods: Insights from the Field.* Available at: http://surveyinsights.org/?p=2991.

Torres van Grinsven, V., Bolko, I., and Bavdaž, M. 2014. "In Search of Motivation for the Business Survey Response Task." *Journal of Official Statistics* 30(4): 579–606. Available at: http://www.degruyter.com/view/j/jos.2014.30.issue-4/jos-2014-0039/jos-2014-0039.xml?format=INT.

Wagner, J. 2010. "The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data." *Public Opinion Quarterly* 74: 223–243. DOI: 10.1093/poq/nfq007.

Wagner, J. 2012. "A Comparison of Alternative Indicators for the Risk of Nonresponse Bias." *Public Opinion Quarterly* 76 (3): 555-575. DOI: 10.1093/poq/nfs032.

Willimack, D. and Nichols, E. 2010. "A Hybrid Response Process Model for Business Surveys." *Journal of Official Statistics* 26: 3-24. Available at: http://www.jos.nu/Articles/abstract.asp?article=26000

Appendix

Negative Flyer Mailed with 2nd NRFU Letter

UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau

2014 Annual Survey of Manufactures

Your response is overdue:

Your MA-10000 report was due February 12, 2015.

Within the next 10 days, please visit us online to begin reporting. Refer to the enclosed letter for our secure website and your log in credentials.
If you cannot report at this time, please go online and request an extension so we know when to expect your report.

Your participation is required by law:

This survey is mandatory under an Act of Congress. Title 13, United States Code, Sections 182 and 224, requires your response. Section 9 guarantees that your response is confidential and will be used for statistical purposes only.

Your data are critical:

Your manufacturing business is part of a scientifically selected sample and represents other business in the country of similar size and industry. Therefore, it is critical to receive your data to ensure we can publish reliable national estimates on current U.S. manufacturing industry outputs, inputs, and operating status.

SAMPLE

Second Nonresponse Follow-up Letter

UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

This is an official request from the United States Government. The Office of Management and Budget (OMB) approval number for these collections are 06070444, which expires 9/30/2016 and 0607-0449, which expires 4/30/2015.

A Message from the Director, U.S. Census Bureau:

Your 2014 Report of Organization, NC-99001, and 2014 Annual Survey of Manufactures, MA-10000(L) are **overdue**. These surveys are **mandatory** and were **due by February 12, 2015**.

Please report within 10 days. To begin, go to the website noted below, click the "Report Now" button, and enter your user ID and password.

Website:
User ID:
Password:

If you recently submitted your surveys, please visit the website noted above and use the "Self Service Log in" to verify your filing status.

Please Note: The U.S. Census Bureau is moving toward all-electronic collections. In addition to reducing our costs, the benefits to reporting electronically include:

- Built-in calculation features
- Edits that identify potential reporting issues
- Immediate confirmation of submission

The Report of Organization is used to update the Census Bureau's list of businesses, and it provides key source data for the County Business Patterns program and other statistical series. Companies, business analysts, and trade associations use this information in planning investments, production, and marketing.

The Annual Survey of Manufactures collects key data about the manufacturing sector of the nation's economy. These data are used to develop accurate estimates of domestic output and productivity, and they serve as a valuable resource for making sound decisions on economic trade policies.

Title 13, United States Code, Sections 182 and 224, **requires your response** and Section 9 guarantees that **your response is confidential** and will be used for statistical purposes only. Applicable provisions of the law are provided on the website noted above. You may use reasonable estimates if book figures are not readily available.

If you need assistance beyond what is available on our website, please contact us at 1-800-233-6136 Monday through Friday, 8:00 a.m. to 4:30 p.m. Eastern Time.

Thank you for your cooperation.

United States Census Bureau
Economic Statistics

SAMPLE