Identifying Establishment Characteristics of Potential Nonresponse Bias

Morgan Earp[1], Daniell Toth[1], Polly Phipps[1], & Charlotte Oslund[1]

[1]Bureau of Labor Statistics, PSB Suite 5930, 2 Massachusetts Avenue NE,
Washington, DC 20212

Abstract

This paper attempts to identify establishment characteristic associated with lower response rates and nonresponse bias in the Job Openings and Labor Turnover Survey of the U.S. Bureau of Labor Statistics. Using regression trees, we identify subgroups of establishments, based on establishment characteristics, that are least likely to respond at each phase of the data collection process, as well as those that contribute to overall nonresponse bias. The results of our regression tree models can be used to develop strategies for increasing participation using responsive design and/or improving post adjustment methods such as weighting.

Key Words: Bureau of Labor Statistics; establishment survey; Job Openings and Labor Turnover Survey; regression trees; response rates; longitudinal survey

1. Introduction

The Bureau of Labor Statistics' (BLS) Job Openings and Labor Turnover Survey (JOLTS) collects data every month from establishments to provide national estimates of job openings, hires, and total separations in the United States. JOLTS samples approximately 16,000 establishments per month from all 50 states and includes both the government and private sectors. JOLTS is a longitudinal survey; once selected, an establishment remains in the survey for 24 months. A new panel of establishments is rotated in and out every month. The JOLTS selects establishments for each panel using a sample stratified by ownership (private or public), region, industry sector, and employment size class.

JOLTS attempts to maintain high response rates since low response rates carry the threat of nonresponse bias, loss of stakeholder confidence, and the potential to inflate variance in survey estimates. Maintaining high response rates requires substantial effort and resources. Traditionally, survey methodologists have used several approaches to increase response rates, such as providing incentives to responders, notification letters, increasing contact attempts, or providing alternative data collection modes (Dillman, 1978; Dillman, Smyth, and Christian 2009; Groves et al., 2002).

Since JOLTS is a longitudinal survey, JOLTS attempts to locate and enroll potential respondents prior to data collection. Each establishment sampled in JOLTS goes through 1) address refinement -- where the address is verified; 2) enrollment -- where the establishment is recruited to participate in the survey; and 3) data collection (see Figure 1). An establishment can become a nonresponding unit during any one of these phases. By modeling each phase separately we can identify establishment characteristics associated with low response rates during each phase so that the BLS can make the best use of resources throughout the data collection process.

Using regression trees, we identify subgroups of establishments least likely to respond at each phase, as well as those that contribute to overall nonresponse bias. The results of our regression tree models can be used to develop strategies for increasing participation such as responsive design and/or improving post adjustment methods such as weighting (Phipps and Toth 2012).

Section 2 contains the methodology and results from modelling the response rates at each phase of data collection for the 2012 JOLTS survey. We provide an interpretation of the separate regression tree models for each phase. In Section 3 we present an analysis of potential nonresponse bias for two key survey variables using a regression tree model and proxy variables created from administrative data. Section 4 contains a discussion of the main results of the analysis.

2. Identifying Characteristics of Nonresponse at Each Phase of Data Collection

In this section we use regression tree models to identify establishment characteristics associated with low response for each phase of JOLTS data collection. The regression tree models use establishment characteristic variables contained in auxiliary data from the BLS Quarterly Census of Employment and Wages, which is the sample frame for JOLTS. The variables we consider for the analyses are shown in Table 1.

Table 1: Establishment Characteristics

Variables	**Description**
Ownership	Federal, State, or Local government, or Private
Private	Government or Private
Sector	Industry Sectors (20)
Super Sector	Groupings of Industry Sectors (11)
White Collar Services	Information, Professional Businesses, & Financial Services vs. All Other Super Sectors
Certainty Unit	Certainty Unit or Not Certainty Unit (certainty units are large/influential establishments selected into the sample with certainty)
Multi-Establishment	Multi-Establishment business or a Non Multi-Establishment business
Size Class	1-9, 10-49, 50-249, 250-999, 1,000-4,999, or 5,000+ Employees
Time in Survey	Number of Months in Survey (0 = First Month in Survey)

Using regression trees, we recursively split the data based on the auxiliary characteristic variables and JOLTS response propensity. At each iteration the variable and breakpoint are chosen to maximize the heterogeneity across subgroups and the homogeneity within groups with regard to nonresponse. The regression tree models presented are pruned versions of trees built using the CRT (Classification Regression Tree) method in SPSS. The pruning is done to provide easily interpretable relationships between establishment characteristics and their impact on response rates.

Our study sample consisted of 207,567establishments sampled for JOLTS during 2012. We excluded establishments that were out of business (*n* = 12,734) and post offices (*n* = 480). Post offices were excluded as the postal service provides data to JOLTS as a census by state. We excluded a small number of establishments with no record of any contact or collection attempt since we are interested in classifying establishments that do not respond given the opportunity. After removing these records, our final dataset used for analysis consisted of 194,353 establishments.

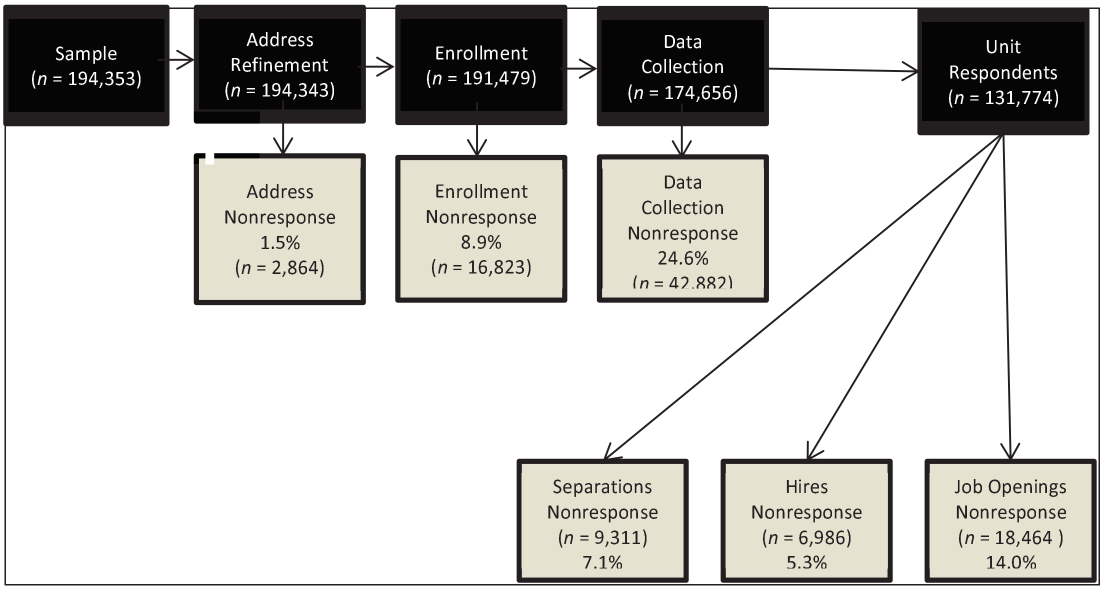
Sample (n = 194,353)
Address Refinement (n = 194,343)
Enrollment (n = 191,479)
Data Collection (n = 174,656)
Unit Respondents (n = 131,774)
Address Nonresponse 1.5% (n = 2,864)
Enrollment Nonresponse 8.9% (n = 16,823)
Data Collection Nonresponse 24.6% (n = 42,882)
Separations Nonresponse (n = 9,311) 7.1%
Hires Nonresponse (n = 6,986) 5.3%
Job Openings Nonresponse (n = 18,464) 14.0%

Figure 1: Jolts Data Collection Phases

2.1 Address Refinement

During address refinement, BLS locates and verifies the contact information of sampled establishments by telephone. Establishment contact information is provided by states and is included as part of the sample frame. By the time the frame is used to draw the sample for JOLTS this contact information is at least 12 months old. Most sampled establishments have some known contact information, but there are a few with little or no contact information available. Even in the case where contact information is provided, the quality and extent varies. A street address is provided for most establishments, and in some cases a telephone number, but for the majority there is no contact name. Also, when contact information is available, it may be out of date, given the 12-month lag time. If the contact information for an establishment cannot be verified by the BLS, these establishments are considered nonrespondents during the address refinement phase (BLS, 2013a).

At first glance it seems that response is high for address refinement, 98.5 percent, and therefore may not be of much concern, as shown in the top box of the regression tree in Figure 2. However, the tree model, identifies two groups with significantly lower response rates. The first group is federal government, with a response rate of 86.9 percent (Node 1); 11.6 percentage points below the overall response rate at this phase. This is likely due to the fact that a) it can be hard to find the correct building for large federal agencies; and b) that the list frame federal agencies are sampled from is often missing the address and specific location of the establishment. The second group with lower response rates during address refinement includes establishments in the retail trade sector with 250 or more employees, that are part of multi-establishment firms, with a response rate of 47.3 percent (Node 8); 51.2 percentage points below the overall response rate at this phase.

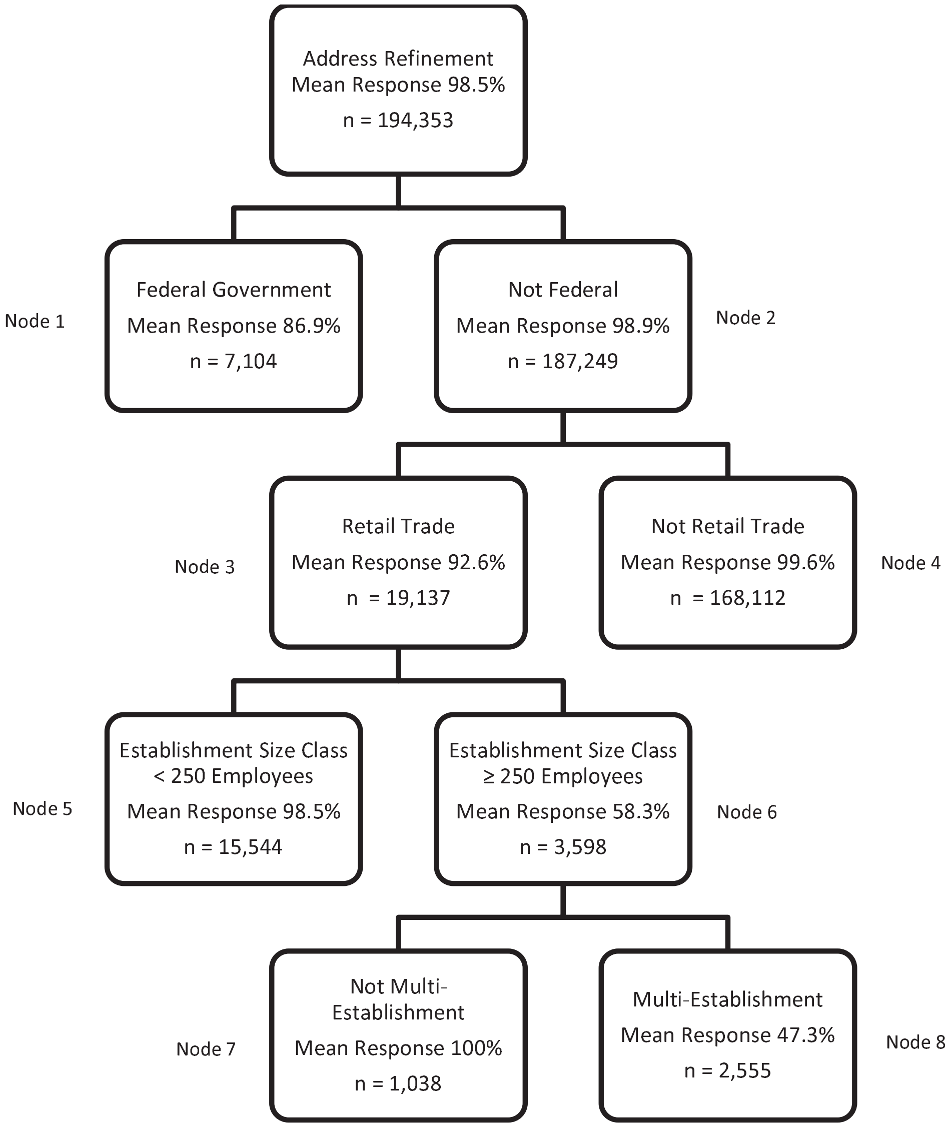

Figure 2: Address Refinement Response Tree Model

2.2 Enrollment

Once an establishment's address is verified, it moves into the enrollment phase; if an establishment's address is not verified then it does not continue into the subsequent enrollment and data collection phases (see Figure 1). Enrollment is the second recruitment phase of JOLTS, where interviewers contact establishments and solicit their participation in the survey. The goal of the enrollment phase is to gain consent from the establishment to participate in the JOLTS program, which involves providing monthly employment and turnover data. During the enrollment phase, each establishment is mailed an "introductory packet" explaining the survey and the importance of their participation; these packets include a customized cover letter, JOLTS Brochure, Business Information Guide, Fact Sheet explaining how the data are used, and a JOLTS Survey Form. About three to five days after the introductory packet is mailed out, interviewers follow-up by calling the establishment to solicit participation (BLS, 2013b).

Enrollment response is modeled only for establishments that responded during the address refinement phase; therefore, the response rates modeled during the enrollment phase are conditional on establishments responding during the address refinement phase. The response rate during enrollment (90.9 %) is lower than address refinement (Figure 3). The model identifies two groups with significantly lower enrollment response rates. The first group includes establishments that are part of multi-establishment firms, privately owned, and in the white collar service industries (Node 11), with a response rate of 81.4 percent; 9.5 percentage points below the overall response rate at this phase. The second group includes establishments that are not part of a multi-establishment firm, with 250 employees or more, and in white collar services, with a response rate of 77.9 percent (Node 9); 13 percentage points below the overall response rate at this phase.

Similar to address refinement, the enrollment model exhibits a significant relationship between establishment ownership type and industry; in this case, private ownership and white-collar service sectors lowered response. Also similar to address refinement, the enrollment model shows that being part of a multi-establishment firm is linked to lower response rates. Both models also exhibit a relationship between larger establishment sizes (≥ 250 employees) and lower response rates. The association between large and multi establishments with lower response rates is typical in establishment surveys, while establishments in the white-collar services industry sectors have been shown to have low response rates compared to other federal surveys (Phipps and Toth 2012).

2.3 Data Collection

After an establishment is successfully enrolled in the survey, the interviewer schedules an appointment and moves the unit into the data collection phase, at which point, the interviewer attempts to collect the requested data. Establishments are asked every month to report the number of employees, hires, total separations, and job openings over a 24-month period, except for certainty units which remain in the survey indefinitely. For the first five months, most establishments complete the survey via computer-assisted telephone interviewing (CATI); after that time, an establishment may be transitioned to other data collection modes like Web, Email, or fax. Offering a variety of collection methods helps accommodate respondent preferences, which is important since JOLTS is a voluntary survey program (BLS, 2012).

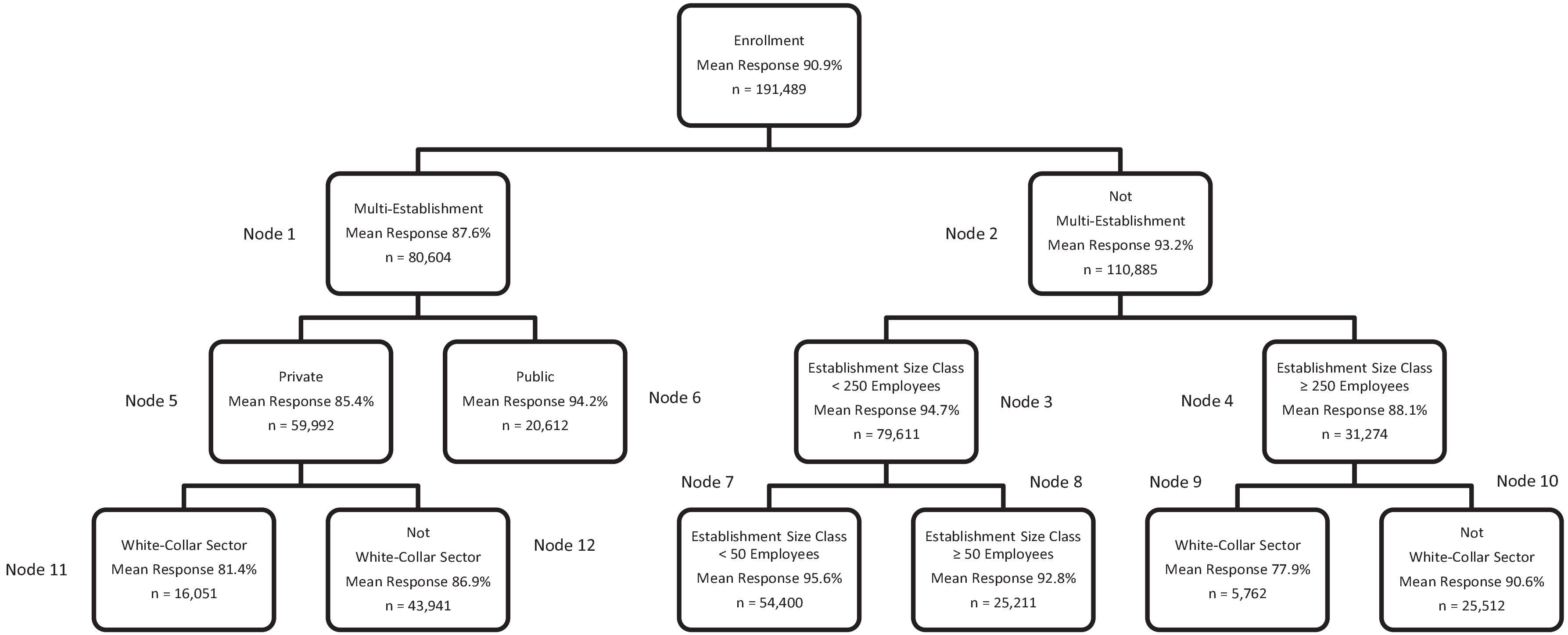

Figure 3: Enrollment Response Tree Model

In order to be counted as a respondent in the data collection phase, the establishment must have provided data that was used for the survey estimates; therefore, the response rates modeled during the data collection phase are conditional on establishments responding during the enrollment phase. Since the data collection phase is different than the address refinement and enrollment phase in that it typically lasts for months, we had to weight the observations to account for the multiple times an establishment appears in the data, since some establishments may have appeared up to 12 times in the dataset.

Figure 4 shows there is an overall 76.9 percent response rate during the data collection phase. Just like in the address refinement and enrollment models, we see a relationship between establishment ownership type, industry type, being part of a multi-establishment firm, and size. Something that is unique to the data collection model is the added negative effect of being a certainty unit and amount of time in the survey. According to Figure 4, privately owned certainty units with 250 or more employees only have a 54.3 percent response rate (Node 8); 22.6 percentage points below the overall response rate for this phase. It is even lower for those in their first month of data collection at 15.3 percent (Node 11); 61.6 percentage points below the overall response rate for this phase.

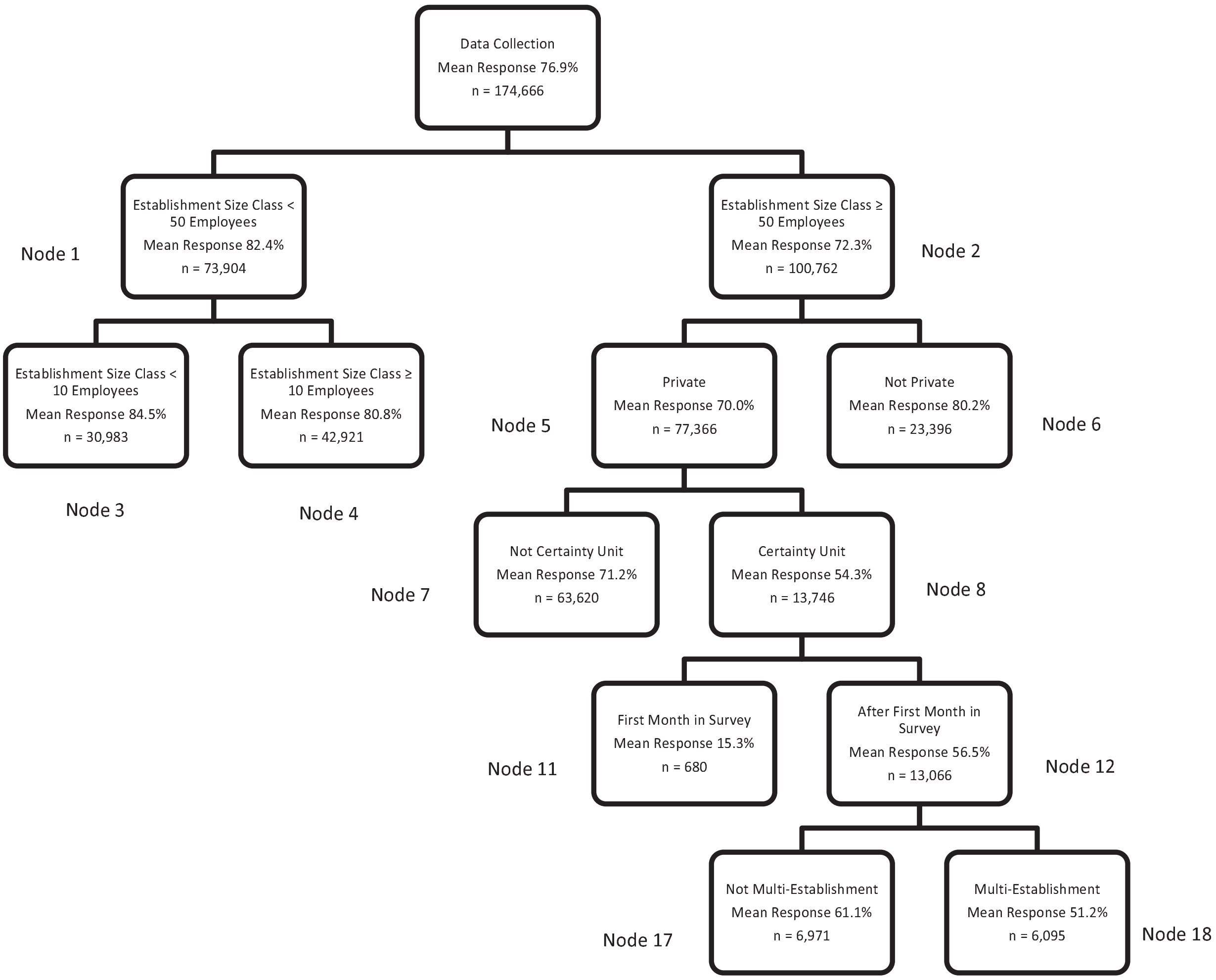

Figure 4: Data Collection Nonresponse Tree Model[1]

[1] Nodes 9, 10, 13, 14, and 15 were pruned from the model, since they did not exhibit large differences in response rates.

2.4 Summary

By looking at each phase of nonresponse separately, we can see that the characteristics of nonrespondents vary, which helps us to better understand when and for what types of establishments nonresponse is an issue. For example, we now know that directing efforts toward federal government establishments during data collection would not be nearly as effective as doing so during address refinement. Also, waiting to target white-collar service sector establishments until enrollment is a potential strategy, since there is less difficulty locating and verifying their addresses and contact information compared to gaining their participation in the survey. It is also helpful to know prior to sampling and initial contacts, that employment size and structure (being part of multi-establishment firm) are negatively correlated with response at each phase.

3. Assessing Nonresponse Bias in the Overall Sample

Each of the models provide useful information that can be used by JOLTS to allocate collection efforts to identified groups of establishments at each stage of the data collection process. However, what they do not tell us is where we are likely to see nonresponse error. Table 2 summarizes the tree model for characteristics of JOLTS response for the entire sample. The tree model resulted in twelve end nodes (12 mutually exclusive groups with varying response propensities). Within each of these end nodes, we compare levels of QCEW employment change for nonrespondents and respondents. The QCEW employment change has been grouped into quartiles by size class with a median of zero; the first quartile containing negative change, which indicates high separations, and the fourth quartile containing positive change, which indicates high hires. By comparing the percentage of respondents to nonrespondents with high hires and/or high separations, we are able to explore which types of establishments have potential for nonresponse error after controlling establishment characteristics contained in the model.

The overall sample response model is similar to the address refinement, enrollment, and data collection models in that response rates continue to be linked to establishment size, ownership type, industry type, and structure. Just like during the enrollment and data collection phase, larger establishments consistently have lower response rates, as do privately owned establishments, and establishments that are part of multi-establishment firms. Using a chi-square test for independence to test the association between response status and high hires/separations status (shown in Table 2), we determined that five of 22 subgroups (nodes) had a significantly lower proportion of respondents than nonrespondents classified as high hires (indicating potential to under represent hires); eight had a significantly higher proportion of respondents classified as high hires (indicating potential to over represent hires); 17 had a significant lower proportion of respondents classified as high separations (indicating a potential to under represent separations); and 0 had a significantly higher proportion of respondents classified as high separations (indicating potential to over represent separations).

While we found that 13 of the 22 groups exhibited a significant relationship between response status and high hires, using the binary phi-coefficient of correlation we determined that only one correlation was high enough to even be considered a small effect size according to standards contained in Cohen's book on *Statistical Power Analysis for the Behavioral Sciences* (1998). In addition, while 17 of the 22 groups exhibited a significant relationship between response status and high separations, none of the correlations were high enough to even be considered a small effect size. The one group that exhibited a small negative effect size between response status and high hires, consisted of privately owned white collar service sector establishments with 5,000 or more employees (Node 17), this subgroup has the lowest response rate of any subgroup at 29.5 percent, but they also only make up 0.27 percent of the entire sample. Tracing back this group through the different phases of data collection, it appears they have a 99.6 percent chance of making it through address refinement (Figure 2, Node 4), a 81.4 percent chance of making it through enrollment conditional on responding during address refinement (Figure 3, Node 11), and a 54.3 percent chance of making it through data collection conditional on responding during enrollment (Figure 4, Node 8). Therefore using a responsive design model, it would be best to target the group starting at enrollment, but primarily during the data collection phase, since this is where we tend to be losing them the most. Within Node 8 of the data collection model, we see that these establishments are spread across subsequent Nodes 11, 17, 18; the majority coming from Node 17 (48.9%) and Node 18 (50.2%). These are the privately owned certainty units with one or more months in the survey; so it appears all but less than one percent make it through their first month in data collection, but are lost sometime thereafter. In addition to highlighting what groups may contribute to nonresponse bias, our model of sample nonresponse does a good job of adjusting for nonresponse and minimizing bias in hires and separations (which is why there is little bias leftover), and thus the inverse of the resulting propensity scores could serve as a potential nonresponse weight adjustment.

4. Discussion

This study compares the characteristics of nonresponding establishments across the various phases of data collection – both before and during data collection. At all phases, we found that employment size and structure (being part of a multi-establishment firm) was negatively correlated with response rates. We also saw that during all three phases, ownership status was correlated with response rates; during address refinement there was a negative correlation associated with federal government, and during enrollment and data collection there was a negative correlation associated with private ownership. The type of service sector also had significant effects in address refinement (we saw negative correlations for retail trade) and enrollment (we saw negative correlations with white collar sector services). Our findings on higher nonresponse rates for larger employment size, white-collar sector services, and multi-establishment firms are similar to those observed in another BLS survey, the Occupational Employment Statistics survey.

Understanding the phases of nonresponse is important in understanding for what types of establishment nonresponse is an issue, but linking that knowledge with an assessment of potential bias provides necessary direction in where to concentrate efforts to reduce nonresponse error. For JOLTS, we were able to explore bias for two key data items -- hires and separations-- using the QCEW data to identify establishments with high hires and high separations. While there were several characteristics that came up as being significantly related to nonresponse (groups where we saw significant differences between respondents and nonrespondents in terms of high hires or high separations) there was only one group that exhibited even a small effect size in terms of hires nonresponse bias; Privately owned establishments with 5,000 or more employees in white collar sector services. While other groups showed significant differences in QCEW hires and separations levels between respondent and nonrespondents, the differences were less than 10 percentage points.

By examining survey response at each survey phase, we can better understand which type of establishments are more difficult to locate, to enroll in the survey, and/or to collect data from each month. During address refinement we suggest focusing on federal government establishments and retail trade establishment with 250 or more employees that are part of multi-establishment firms. We further suggest focusing on enrollment and data collection to target white-collar service sector establishments, since they are not specifically a problem during address refinement. Lastly, it is important to note that employment size and structure (being part of multi-establishment firm) are not only negatively correlated with response at each phase, but are also significantly related to nonresponse bias. Overall our model of sample nonresponse does a good job of adjusting for nonresponse and minimizing bias in hires and separations, and thus the inverse of the resulting propensity scores could serve as potential nonresponse weight adjustments.

5. References

Bureau of Labor Statistics. (2012). *Job Openings and Labor Turnover Survey Data Collection Training Manual.*

Bureau of Labor Statistics. (2013a). *Job Openings and Labor Turnover Survey Address Refinement Training Manual*.

Bureau of Labor Statistics. (2013b). *Job Openings and Labor Turnover Survey Enrollment Training Manual.*

Cohen, Jacob, 1988, *Statistical power and analysis for the behavioral sciences (2nd ed.),* Hillsdale, N.J., Lawrence Erlbaum Associates, Inc.

Dillman, D. 1978. *Mail and Telephone Surveys: The Total Design Method*. New York: Wiley & Sons.

Dillman, D, J. Smyth, and L. Christian. 2009. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method.* New Jersey: John Wiley & Sons.

Earp, M., Toth, D., Phipps, P., and Oslund, C. 2013 Identifying and Comparing Characteristics of Nonrespondents throughout the Data Collection Process. http://www.bls.gov/osmr/pdf/st130090.pdf

Groves, R.M., D. Dillman, J.L. Eltinge, and R. Little (Eds.). 2002. Survey Nonresponse. New York: Wiley.

Kalton, G. & Flores-Cervantes, I. (2003). Weighting Methods. *Journal of Official Statistics*, 19 (2), 81-97.

Phipps, P. & Toth, D. (2012). Analyzing Establishment Nonresponse Using and Interpretable Regression Tree Model with Linked Administrative Data. *Annals of Applied Statistics, 6* (2), 772-794.

Table 2: Nonresponse Bias Tree Results by Establishment Characteristics and Size Classes

Node	Size Class	Private	White Collar Service Sector	Multi-Establishment	Number of Observations	Response Rate	Percent High Hires / Respondents	Percent High Hires / Non-respondents	Hires NR Bias Effect Size φ	Percent High Separations / Respondents	Percent High Separations / Non-respondents	Separations NR Bias Effect Size φ
7	1-9			No	25,940	81.7	7.75	7.29	+.01	30.66	36.63	**-.05*** **
8	10-49			No	28,603	75.9	10.65	12.47	**-.02*** **	31.81	37.14	**-.05*** **
9	1-9	Yes		Yes	3,889	65.3	8.33	7.09	**+.02***	32.57	41.46	**-.09*** **
9	10-49	Yes		Yes	13,876	63.4	9.50	9.67	-.003	30.04	34.07	**-.04*** **
16	50-249	Yes		Yes	19,208	52.5	15.65	16.50	**-.01** **	36.09	41.13	**-.05*** **
10	1-9	No		Yes	2,849	77.7	7.33	5.65	**+.03***	28.75	33.06	**-.04*** **
10	10-49	No		Yes	3,313	75.2	8.98	8.64	+.01	26.23	34.66	**-.08*** **
13	50-249	No		No	7,842	72.4	16.42	14.14	**+.03*** **	33.07	37.56	**-.04*** **
13	250-999	No		No	6,299	70.7	19.00	16.79	**+.02** **	35.83	39.67	**-.03*** **
14	1,000-4,999	No		No	6,915	62.6	21.10	16.01	**+.06*** **	38.78	39.90	-.01
14	5,000+	No		No	4,224	58.1	24.88	23.94	+.01	33.33	34.43	-.01
17	250-999	Yes	Yes		7,559	38.7	20.14	18.35	**+.02** **	43.57	47.89	**-.04*** **
17	1,000-4,999	Yes	Yes		4,537	28.9	19.02	19.01	+.0001	39.40	46.03	**-.07*** **
17	5,000+	Yes	Yes		524	29.5	14.05	22.29	**-.10***	49.59	47.13	+.02
19	50-249	Yes	Yes	No	5,885	55.0	16.85	17.28	-.01	38.21	42.81	**-.05*** **
20	50-249	Yes	No	No	17,312	66.7	17.66	18.90	**-.01** **	37.59	41.23	**-.03*** **
21	250-999	Yes	No	No	9,135	57.7	19.90	19.53	+.004	39.32	44.45	**-.05*** **
21	1,000-4,999	Yes	No	No	7,628	49.6	21.49	19.85	**+.02***	41.35	43.23	**-.02***
21	5,000+	Yes	No	No	1,017	43.5	17.32	22.71	**-.07***	26.26	30.63	-.05
22	250-999	Yes	No	Yes	8,649	42.7	17.91	17.46	+.01	39.72	40.87	-.01
22	1,000-4,999	Yes	No	Yes	7,334	48.4	19.19	17.81	**+.02***	38.14	41.07	**-.03*** **
22	5,000+	Yes	No	Yes	1,815	47.6	17.98	18.99	-.01	28.27	33.76	**-.06***

* Significant at the .05 level; ** Significant at the .01 level; *** Significant at the .001 level

- Effect Sizes (Cohen, 1988): 10-29 Small; 30-49 Medium; 50+ Large