The O*NET Data Collection Program: Improving Efficiency in a Multistage Complex Establishment Survey

Marcus Berzofsky[1], Brandon Welch[1], Susan McRitchie[1], and Rick Williams[1]
RTI International[1]

Abstract

The O*NET™ Data Collection Program, which began data collection in June 2001, is a very large probability-based business establishment survey, with over 143,000 business establishments sampled to date. The study is sponsored by the U.S. Department of Labor and is conducted by the National Center for O*NET Development and RTI International. The survey derives national estimates for 810 occupations across four domains: skills, work content, work activities, and knowledge. Therefore, this study is simultaneously collecting data from 3,240 distinct surveys, resulting in many interesting survey sampling issues. This paper describes the sample design currently being used as well as its evolution over the past 5 years, which has resulted in a more efficient design. We describe the process for identifying the industries in which an occupation is located and how to maximize the overlap between occupations selected at a business establishment while minimizing business establishment burden, the coverage requirements that are needed to ensure a representative sample while simultaneously minimizing cost, the composite size measure that is used to ensure that establishments with the highest likelihood of employing an occupation of interest will also have the largest probability of selection, and our decision to use the Dun & Bradstreet business list for our frame and the benefits associated with that choice. Then, we describe how these factors are used to inform the two-stage process of first drawing a sample of business establishments and then drawing a sample of employees in the occupations in the targeted occupations. Next, we describe the wave design that allowed us to efficiently control the sample size and target industries for each occupation. Finally, we discuss model-aided sampling (MAS), which adds additional procedures to control the sample allocation.

Keywords: O*NET, business establishment survey, coverage, burden, composite size measure, Dun & Bradstreet, wave design, model-aided sampling.

1. Introduction

Sponsored by the U.S. Department of Labor and conducted by the National Center for O*NET Development and RTI International, the O*NET Data Collection Program provides a database containing information on a multitude of occupational attributes. Information related to these attributes is acquired using a nationally representative business establishment survey. The survey is designed based on a deductive approach in which a common set of prespecified items are defined and data are collected on those items. Data are collected and estimates are produced for four occupational domains: skills, work content, work activities, and knowledge. Therefore, the O*NET Data Collection Program is simultaneously collecting data on 3,240 surveys. The current program goal is to produce estimates on 810 O*NET occupations (plus new and emerging occupations).

The target population for the O*NET Data Collection Program is defined as all nonmilitary, noninstitutionalized job incumbents working in the 810 O*NET occupations within the 50 United States plus the District of Columbia. To date, over 143,000 establishments have been selected, and over 100,000 incumbent employees have responded. The main source of data collection is a traditional two-stage sampling paradigm; however, for occupations where this model is not efficient, other list-based methods are used. This paper focuses solely on the traditional design. As of 2007, estimates have been produced for more than 700 of the 810 aforesaid occupations. Each year the O*NET database is updated with new information.

The primary goal of this paper is to outline some of the challenges faced since the project's inception and to discuss some of the solutions implemented. We begin by discussing some of the overall design challenges faced when identifying incumbents in all 810 occupations. Next, we describe the initial sampling design and how the wave design has evolved. Finally,

we describe the introduction of model-aided sampling (MAS [1]) and how it is used in data collection.

2. Design Challenges

The O*NET Data Collection Program posed several design challenges that needed to be resolved prior to data collection. First, a method for identifying incumbents in each of the 810 occupations needed to be determined. Ideally, a single method of sample selection would be used for all of the occupations. Since we were interested in producing estimates for all 810 occupations, we needed a method to identify the population of each occupation. This is different than most large-scale surveys where the primary interest is in producing estimates at the national level and for a few key subpopulations.

A simple solution to this problem is to generate a list of incumbents in each occupation. Then, a simple random sample of incumbents is drawn, and data are collected in a relatively straightforward manner. For example, one could obtain a list of all licensed lawyers from the American Bar Association or a list of doctors by specialty from the American Medical Association. For occupations such as these, the list approach seems very inviting. Unfortunately, it is not possible to obtain a list for all occupations. For instance, there is not a comprehensive list of secretaries in the United States. Therefore, the list approach does not achieve the goal of developing a single methodology that is applied to all occupations.

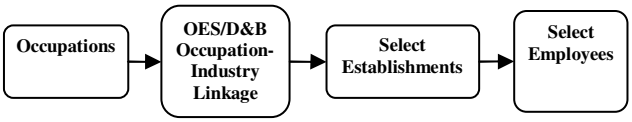
A second more complete solution is to take advantage of the likelihood of occupations belonging to the same industry and, hence, the same establishment. Thus, one could conduct a general population survey of establishments and then sample incumbents within selected establishments. Since all incumbents are employed at an establishment, this approach allows us to simultaneously sample all desired occupations.

Although the above approach is more complete, it also creates the logistical challenge of identifying a mechanism to fully identify the set of industries, and locate the establishments within those industries, that include the occupations of interest. To overcome these challenges, the O*NET Data Collection Program used the Occupational Employment Statistics (OES) survey, which is conducted by the U.S. Bureau of Labor Statistics, and the Dun & Bradstreet (D&B) frame of business establishments. OES provides two critical pieces of information that resolve the challenge of identifying the industries and estimating the employment. First, OES fully identifies the industries linked to all occupations. This allows us to define the target population of a given industry. Secondly, it obtains estimates on the number of employees for an occupation in each industry. This allows us to determine the magnitude by which an occupation is more likely to be found in one industry compared with another. Information provided by D&B resolves the challenge of identifying establishments since they are able to provide a frame of all establishments by industry and the total number of employees in each of those industries. Furthermore, D&B categorizes establishments by the number of employees at the establishment, which allows better estimation of the likely number of employees in an occupation that will be found in an establishment of a particular size and in a particular industry. Since OES and D&B contain industry information, we are able to combine them by industry. Therefore, we can determine the industries containing particular occupations and develop a mechanism to select establishments in those industries.

3. Initial Design

After overcoming these early design challenges, we developed the initial design used to sample incumbents in the occupations of interest. An initial sample size of 12,000 establishments was allocated to target 210 occupations. These 210 occupations covered a wide range of occupations across all industry types. Sampling was achieved through a traditional probability based two-stage cluster design. The first stage involved selecting establishments and linking occupations to those establishments, and the second stage involved selecting employees from the occupations associated with those establishments. This process is illustrated in Figure 1.

Figure 1. Initial Wave Life Cycle



Prior to the first stage of selection, occupations were linked to industries based on information provided by OES. Based on this industry-occupation association a composite size measure (CSM) (Folsom, Potter, & Williams, 1987) was derived for each establishment in a particular industry based on the occupations linked to the establishment. Industries were stratified into five groups, and the sample of 12,000 establishments was allocated across strata proportional to their size measure. Establishments were selected within strata

[1] In a previous publication, MAS was defined as model-assisted sampling.

with probability proportionate to their CSM using the Sequential PPS method[2] (Chromy, 1979).
After an establishment was selected, PPS random sampling was used to link up to 10 occupations to an establishment. The list sample method is based on a simple random sample design and, therefore, does not allow one to take advantage of the inherent clustering of similar occupations within an establishment. Therefore, the establishment method reduces the cost of data collection compared with the list method. In an effort to reduce establishment burden, the number of occupations linked to a selected establishment was limited to 10. By selecting these 10 via PPS, it was more likely that occupations with a greater likelihood of being present at an establishment were selected.

Once an establishment was selected and contacted, a point of contact (POC) at the establishment was identified. For all occupations present at the establishment, the POC was asked to roster the employees in each occupation. A simple random sample of 15 employees was selected from across all rostered employees. The selected incumbents were then randomized to one of the four questionnaire domains. Questionnaires were coordinated through the POC and could either be completed through a paper form or through an identical questionnaire on the Web.

4. Wave Design

Once the initial sample was drawn and data collection began, several limitations of the initial design were identified. First, it inefficiently covered all occupations. The initial set of occupations was so diverse that they were found in almost all industries. Therefore, even with a sample size of 12,000 establishments, some industries had few or no establishments selected. Thus, some occupations were not adequately targeted, which negatively affected the occupation's coverage. This impacted data collection on two fronts. First, occupations in fringe industries were difficult to locate, which created an inefficient use of the establishments that were selected. Secondly, because the sample was clustered in the larger industry areas, some occupations that are primarily found in smaller industries were linked to an insufficient number of establishments. These inefficiencies in the initial sample design limited the information available to inform future follow-up samples that were necessary to complete the initial set of occupations.

[2] For the initial design, the sampling frame was constructed using InfoUSA's list of business establishments (not D&B).

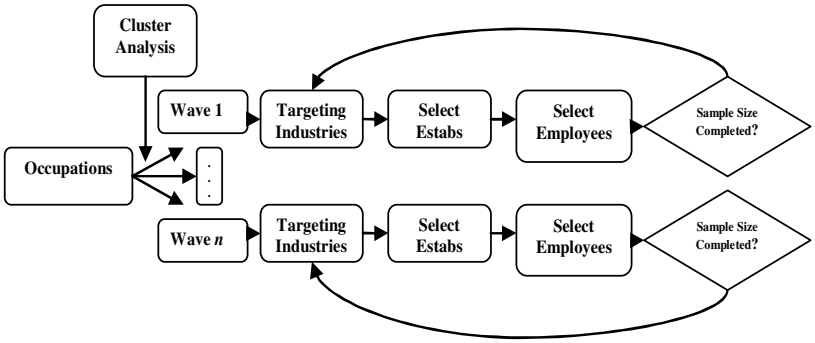
A wave design was developed in response to the limitations encountered in the initial sample. The wave design modified the sample design in three ways. First, it split the occupations into smaller groups. Second, in the wave design we determined that the coverage level for an occupation could be relaxed without introducing significant bias and targeted industries more efficiently to better achieve a high yield sample. Finally, we released the sample for these smaller sets of occupations in subwaves of 3,000 establishments. The wave design is illustrated in Figure 2.

The first major change of the wave design was that we split the set of 810 occupations into groups of approximately 50 occupations. These groups are denoted by *X.y*, where *X* is the sampling wave and *y* is the sampling subwave. All sampling waves were initially denoted by *X.1*. After 6 months, we assessed the number of completed questionnaires, and those lacking 15 completed questionnaires in each domain were considered for inclusion in the next sampling subwave, *X.2*. The typical life cycle of a wave was *X.4*.

To determine which occupations belonged to a particular wave, we used a cluster analysis to group the occupations based on their distribution of industries according to OES. Industries were excluded from the analysis if 1 percent or less of the employees in an occupation, for all occupations, were found in that industry. Each cluster contained 3 to 20 occupations, and these clusters were used to form the waves containing 50 occupations. Clustering the occupations that were found in a common set of industries and then forming waves should increase the likelihood of finding all of the occupations in the set.

The second modification made in the wave design was reducing the required coverage level for an occupation and targeting industries for an occupation based on the likelihood of finding the occupation in a particular industry. During the initial design, a high coverage level of employees in an occupation was required for each occupation, as is typical under a traditional establishment sample design. However, as noted, this led to inefficiencies by linking occupations to establishments for which there was only a small chance of being present. To see if these coverage requirements could be relaxed, we conducted a coverage analysis that compared the estimates from respondents in industries for which an occupation is typically found with more fringe industries for the occupation. The results of this analysis showed that the difference in the estimates between these two industry groupings was not statistically significant across a wide set of occupations. Therefore, under the wave

Figure 2. Wave Design Life Cycle



design, an occupation was not linked to the more fringe industries identified by OES. We define a fringe industry for an occupation as an industry containing a small percentage of an occupation. For example, brick layers may work at hospitals, but it would be inefficient to sample hospitals in hopes of finding brick layers.

Furthermore, further targeting was executed to increase the likelihood of selecting industries linked to the occupations of interest. To achieve this goal, substantive experts, Industrial Organization (I/O) psychologists, assigned a concentration level to each of the industries for which an occupation was linked. These concentration levels were used to adjust the CSM, thus, improving the likelihood of selection.

The next major modification of the wave design was to release the sample in smaller subwaves. For a particular set of 50 occupations, a sample of 3,000 establishments was drawn. This initial subwave was similar to the initial design in that it solely relied on OES information. After data collection for the initial subwave was conducted, a second subwave was designed. This second subwave combined OES information with the information obtained during the initial subwave. Substantive experts used this empirical information to assign concentration levels to the industries associated with each occupation. The subwaves continued until all occupations met the minimum requirement for number of completed questionnaires. Furthermore, by splitting the design into subwaves, we were able to introduce additional industry stratification in later subwaves that ensured that all occupations were linked to a fair number of establishments. This allowed us to better allocate the sample to harder-to-locate occupations.

In addition to making changes to the manner in which establishments were selected, we modified the manner in which survey protocol burdened study participants. Two types of burden were considered: establishment burden and employee burden. Establishment burden applies when the POC determines if the occupations are present at the establishment, rosters employees, and distributes questionnaires to selected employees. Time used by the incumbent to complete the questionnaire is considered employee burden.

We made several changes to our establishment sampling protocol to minimize burden. Inquiries regarding the 10 occupations linked to the establishment ceased after 5 occupations were found present. This reduced the number of occupations the POC rostered to no more than 5 occupations. Furthermore, once an establishment was selected we deemed it ineligible for reselection for the next 12 months. This rule held even if we targeted a different set of occupations. Moreover, we modified our employee selection algorithm so that no more than 20 employees were selected from an establishment and no more than 8 employees were selected from one occupation. Therefore, if only one occupation was present at an establishment, no more than 8 employees would be selected. This change benefited the establishment in that we were not overburdening a particular occupation at the establishment. In addition, it helped statistically in that it limited the number of respondents for an occupation from one single location, which reduced the cluster effect one establishment had on an occupation's estimates.

5. Model-Aided Sampling

Under the wave design, the amount of effort in collecting information varied by occupation. In some instances, more questionnaires were completed than needed to satisfy the criterion for completeness (15 per domain). Likewise, some occupations required more effort than others to complete. For example, the amount of effort in finding secondary school teachers was minimal, whereas roustabouts were very difficult

to find and responded at low levels. Hence, a new strategy was introduced to minimize the level of effort while ensuring a representative sample.

To achieve a more balanced sample allocation, the O*NET Data Collection Program developed MAS, which combines the traditional sampling approach with quota sampling (Berzofsky, Welch, Williams, & Biemer, 2006). Under the MAS design, establishments are selected using the two-stage sampling techniques used in the wave design, but restrictions are placed on the number of questionnaires needed to complete each occupation. For each occupation, quotas were based on OES information. Within each quota class, we proportionally allocated the quota based on the distribution of employment reported by OES. Exhibit 1 lists the set of classes and subdomains for which quotas are defined. Similar to targeting industries, substantive experts help determine the size of each quota by reviewing the quota specifications for each occupation. Once quotas are defined in the first subwave, establishment sampling is executed under the traditional two-stage approach. Data collection is ceased in a quota cell once the targeted number of questionnaires is completed. Data collection for the entire occupation is ceased after the minimum target is met for all quota cells. Berzofsky et al. (2006) conducted a simulation study that showed estimates under the traditional paradigm were substantively similar to those using the MAS paradigm.

Exhibit 1. MAS quota classifications

Industry division
Agricultural, Forestry, and Fishing
Wholesale Trade
Mining
Retail Trade
Construction
Finance, Insurance, and Real Estate (FIRE)
Manufacturing
Services
Transportation, Communications, Electric, Gas, and Sanitary Services
Government (Federal, State, and Local)
Census region
Northeast
South
Midwest
West
Number of employees
Unknown, 1–9
10–49
50–249
250 or more

MAS improves the O*NET Data Collection Program's design by ensuring that the respondent sample for an occupation is representative while controlling the burden level used. For example, if a particular occupation meets its quota in a particular domain, say the mining industry division, sampling will cease for that division. Thus, further sampling is unnecessary for that division in future subwaves. Likewise, if the number of questionnaires in a specific division is lacking, substantive experts target specific industries in that division, hence, guiding the occupation toward completion. Thus, under MAS, data collection continues until all quota cells have been completed. Future subwaves only target quota cells that have not been completed. Figure 3 illustrates the life cycle for MAS.

6. Conclusions

Improving efficiency on large-scale surveys offers a myriad of design challenges. Often all of these challenges are not clear at the onset of the design. Therefore, the O*NET Data Collection Program exemplifies the need to constantly evaluate the current design and make improvements as necessary. The evolution of the O*NET Data Collection Program design illustrates how relatively small changes in the design can greatly improve efficiency. However, it is critical that any design change be tested before implementation. It is important to know the impact that a change will have on the study, and, thus, survey results, before being implemented. The evolution of the O*NET Data Collection Program design demonstrates how a multiyear study often needs modifications and how, with proper implementation, those modifications can greatly improve the study.

References

Berzofsky, M. E., Welch, B. L., Williams, R. L., & Biemer, P. P. (2006). Using a model-assisted sampling paradigm instead of a traditional sampling paradigm in a nationally representative establishment survey. *Proceedings of the American Statistical Association, Section on Survey Research Methods* (pp. 2763-2770).

Chromy, J. R. (1979). Sequential sample selection methods. *Proceedings of the American Statistical Association, Section on Survey Research Methods* (pp. 401-406).

Folsom, R. E., Potter, F. J., & Williams, S. R. (1987). Notes on a composite size measure for self-weighting samples in multiple domains. *Proceedings of the American Statistical Association, Section on Survey Research Methods* (pp. 792-796).

Figure 3. MAS Design Life Cycle