5/28/97

STATISTICAL PROBLEMS IN BLS COMPENSATION SURVEYS WHEN COLLECTED ESTABLISHMENT DATA DIFFERS FROM THE ASSIGNED DATA

Susan R. Black, Lawrence R. Ernst, and Jason Tehonica

Bureau of Labor Statistics
2 Massachusetts Ave. N.E., Room 3160, Washington, DC 20212

1. Introduction

The National Compensation Survey (NCS) is a Bureau of Labor Statistics (BLS) establishment survey program of employee salaries, wages, and benefits. The program is designed to produce data at local levels, broad regions, and nationwide. The NCS will replace three existing BLS survey programs: Employment Cost Index (ECI), Occupational Compensation Survey (OCS) Program and Employee Benefits Survey (EBS). The NCS was developed to expand the data products of the existing compensation programs to eliminate duplicate data collection and processing requirements, reduce respondent burden, develop more efficient and streamlined collection and processing techniques, and to address budget constraints. Cohen (1997) presents a more detailed overview of the NCS program.

Within each geographic PSU, there are two stages of sampling. An establishment sample is chosen during the first stage of selection. At the second stage of selection an occupation sample is selected within each establishment. For local estimates, the weight for each employee in a selected defined occupations is obtained by taking the product of the reciprocal of the probability of selecting the establishment, the reciprocal of the probability of selecting the defined occupation within the selected establishment, and nonresponse adjustment factors for establishment and occupation nonresponse. All of the terms used to describe the weight of each employee are presented in more detail in Black, Ernst, and Tehonica (1997), along with a more detailed description of the sample design. Other weighting factors used to compensate for situations where the unit collected differs in some way from the originally assigned unit will be described in this paper.

This paper will address the statistical problems in the NCS when collected establishment data differs from the assigned data. A large survey program, such as NCS, attempts to have clear and precise data collection procedures set. But, there are many data collection problems that occur which are out of the hands of the data collector. When a data collector encounters these problems a weight adjustment can be made to lessen the effect on variances and biases of the estimates.

Section 2 covers establishment data collection issues. The most complex issue arises when a respondent provides data for more locations within a survey area than for what is sampled. Also discussed is subsampling a unit that consists of several physical locations or divisions. Adjustment for the situation when a unit is comprised of several locations and one or more of the locations are considered nonrespondents is also covered. "Birth units" and changes in ownership add to the complexity of the issues in Section 2. Section 3 discusses two data collection issues relating to the occupation sampling. First is the subsampling of employees when the selected defined occupation has a large number of incumbents. Second is the changes such as the merger of sampled establishments that can occur after the initial interview. This can impact the occupational data collection and weighting for subsequent interviews. The occupation selection and weighting issues for establishments which are part of a Central Office Collection (COC) procedure are discussed in Section 4. Finally, in Section 5 we discuss how to modify the variance estimation procedures to reflect the special sampling and weighting required for the situations that were discussed in the previous sections.

2. Establishment Issues

Data collection and weighting for establishments in the NCS can be likened, at least in part, to shooting at a moving target. This is because as a result of issues such as "birth establishments," establishments which change ownership, and the inability or unwillingness of the respondent to report data for precisely the unit that it is desired to collect, obtaining unbiased estimates for the data collected can be a daunting and in part an operationally impossible challenge. This section discusses these issues.

Corresponding to each assigned unit is a set of physical locations from which data is actually collected in the initiation interview We envision that there are six steps in determining this set of physical locations. Some of these steps require modifications in the weights to compensate for changes in the set of locations as we proceed through the steps. Not all the steps are required for each sampled unit. In fact for the vast majority of sampled units we can stop after step 1. Understanding these steps requires being able to distinguish between the following terms: "assigned unit," "modified assigned unit," "subsampled assigned unit," "desired collection unit" and "actual collection unit." Each of these terms is defined in the description of the six steps that we now proceed to present.

1. Determination of the "assigned unit." The frame from which the first stage sample units or "establishments" are selected within each sample geographic area for the NCS is a set of UDB numbers which are in one-to-one correspondence with the set of UI reporting numbers. For the most part, a UDB number represents a physical location for a company. However, sometimes a UDB number represents multiple locations for a company. In either case, an "assigned unit" for each selected UDB number is considered to be the set of physical locations corresponding to the UDB number, based on the report to the UI that was used in the frame construction. The same assigned establishment weight is given to each physical location that is part of an assigned unit, namely the reciprocal of the probability of selection of the corresponding UDB number. This weight is modified by an establishment nonresponse

adjustment procedure which adjusts the assigned weights using weighting class cells formed by assigned employment and assigned industry

2. Determination of the "modified assigned unit." For the most part we seek to collect data for the set of physical locations that constitute the assigned sample unit. There are the following two exceptions to the rule of attempting to collect data from precisely the locations that are part of the original assigned unit. An original sample assigned unit as modified by these exceptions is known as a "modified assigned unit."

A. Any location that is part of the assigned unit but is outside of the sample geographic PSU is excluded from the modified assigned unit.

B. Any location within the sample PSU that reports to the UI under the number corresponding to the sampled UDB number that is in business at the time the data is collected but began business too late to be included the original assigned sampled unit, is included in the modified assigned unit.

The locations described in the latter exception are "birth locations." We have a separate birth sample designed to pick up locations that begin business too late to be in the original sampling frame. However, the birth sample units are obtained by matching an updated sampling frame with the original sampling frame to find UDB numbers that were not on the original sampling frame. If a new location is reporting to the UI under a number that corresponds to an existing UDB number it cannot be picked up by the birth sample procedure. Therefore, we include such locations in the modified assigned unit.

In addition to excluding locations that are outside of the sample PSU, locations that have gone out of business or outside of the industrial scope of the survey are not included in the data collection process. These locations may be identified at various points in the six step process.

The assigned weight described in step 1 for a unit also applies to each location that is part of the modified assigned unit.

3. Subsampling of the modified assigned unit. Occasionally for reasons of respondent or interviewer burden it is necessary to subsample the set of locations that comprise the desired collection unit. Data is then collected from the subsampled locations only. To do this, the current total employment, known as the reported employment, is obtained for each location that comprises the modified assigned unit. A pps subsample of locations is selected, with the reported employment as the measure of size. The original assigned weight for the modified assigned unit, as modified by the establishment nonresponse adjustment factor, is multiplied by the reciprocal of the probability of the location being retained in the subsampling process to obtain the weight for each retained location after this step. We designate the set of locations remaining after this step as the "subsampled assigned unit."

4. Splitting of the subsampled assigned unit into "desired collection units." Sometimes the locations remaining after step 3 must be partitioned or split prior to PSO sampling. Typically this occurs when the assigned unit consists of multiple locations of a company, but the data required to perform the PSO sampling is kept separately at each location. It can also occur if some of the locations after step 3 have been sold to a different company, in which case the data for these locations would have to be collected from the new owners. Note that sometimes elements of the split unit can consist of more than one physical location, for example, when more than one location is sold to a company. In any case, each element of the split unit will be known as a "desired collection unit." The splitting step by itself requires no modification of establishment weights since it does not alter the set of locations which are to contribute to the estimates, just where the collection is to take place. Also note that when subsampling locations as noted in step 3 is required, splitting will always be done. Each element of the split for a subsample must consist of a single physical location. This is because each location of the subsample generally has a different weight after step 3 and by splitting the subsample into single physical locations we avoid the problem of creating a desired collection unit consisting of multiple locations with different weights.

5. Adjustment for nonresponse after splitting. When step 4 is required and some, but not all of the desired collection units resulting from the splitting become nonrespondents, a special nonresponse adjustment factor is used. The weight of each respondent desired collection unit associated with the assigned unit is multiplied by this factor. The numerator of this factor is the weight of the original assigned sample unit after establishment nonresponse adjustment times its assigned employment. The denominator is obtained by multiplying the weight after step 4 for each location that is part of a responding desired collection unit by its reported employment and summing the result over all such locations. Note that because assigned employment is used in the numerator and reported employment in the denominator, it is possible that this computation could result in a factor less than 1, in which case we set the factor to be 1. (Unfortunately, we are forced to use assigned employment in the numerator, since we do not collect reported employment for nonresponding locations, and also must use reported employment in the denominator, since assigned employment does not exist for individual locations when the assigned unit consists of multiple locations.) This nonresponse adjustment for splits has also been called an adjustment for "collected less than assigned. This situation is also one of the cases of the "documentation adjustment factor," which is described in the next step.

The splitting just described is known as "splitting before PSO." There is also another type of splitting known as "splitting after PSO." Typically this occurs when an assigned unit consists of multiple locations and the respondent has the information to allow a single PSO
to be done for the assigned unit, but the wage data must be collected separately at each location. The set of employees in a sampled defined occupation in that case is restricted to the employees in the same location as the selected employee. Splitting after PSO is basically an operational procedure with the only statistical implication occurring if there is nonresponse for some locations. Nonresponse in that situation is considered as occupational nonresponse. Splitting after PSO will not be considered further in this section.

6. Determination of "actual collection units." An "actual collection unit" is the set of locations from which data is actually collected corresponding to each responding desired collection unit. The main reason why an actual collection unit may differ from a desired collection unit is the inability or unwillingness of a respondent to separate company data for locations that are in sample desired collection units from those that are not. More specifically is the case where there are separate UDB numbers for each location in a sample PSU and the respondent will only give us combined data for all of their locations in the sample PSU. This includes both data from sampled and nonsampled locations, where each sampled location corresponds to a separate sample desired collection unit. Another case for which actual collection unit would differ from a desired collection unit would occur when a sample UDB number corresponds to multiple locations and the respondent is unable to exclude locations acquired through change in ownership.

We proceed to describe the general weighting methodology used to handle these types of situations. Note that it is possible for two or more sample desired collection units to correspond to the same actual collection unit. For example, this would be the case if each location of a company corresponded to a separate UDB number and two or more of these UDB numbers were selected with the respondent providing combined data for all locations in the PSU. When there is only a single sample desired collection unit corresponding to an actual collection unit which includes additional locations, the collection situation is referred to as "collected more than assigned." When two or more sample desired collection units are included in an actual collection unit, it is referred to as a "merger." However, the same general weighting methodology is used in both cases to weight an actual collection unit that is a union of desired collection units, and we proceed to describe this methodology. We will assume at first in this description that all desired collection units are respondents, that is neither whole establishment nonresponse adjustment nor the weighting adjustment in step 5 for nonresponse after splitting are needed. We then explain the modifications required when there are nonresponding desired collection units.

Let $Y = \sum_{i=1}^{N} Y_i$ be a parameter of interest, where Y_i is the value for the *i*-th actual collection unit in a population consisting of *N* actual collection units. (We assume conceptually that there is a unique actual collection unit corresponding to each desired collection unit, whether the desired collection unit arises from a sampled assigned unit or only nonsampled assigned units. We of course only know the desired collection units and the corresponding actual collection units that are associated with sampled assigned units.) Let $\hat{Y}_i$ be an unbiased estimator of Y_i, that is $E(\hat{Y}_i) = Y_i$. Let w_i, the weight of the *i*-th actual collection unit, be a random variable which is independent of $\hat{Y}_i$ and which satisfies

$$E(w_i) = 1, \tag{1}$$

and let

$$\hat{Y} = \sum_{i=1}^{N} w_i \hat{Y}_i \ . \tag{2}$$

Then, as observed in Ernst (1989), $\hat{Y}$ is an unbiased estimator of $\hat{Y}$, since

$$E(\hat{Y}) = \sum_{i=1}^{N} E(w_i) E(\hat{Y}_i) = \sum_{i=1}^{N} Y_i = Y \, .$$

We will use a special case of this result as follows. Let N_i denote the number of desired collection units corresponding to the *i*-th actual collection unit. Let W_{ij} denote the common weight associated with each location within the *j*-th desired collection unit of the *i*-th actual collection unit after step 3. That is, W_{ij} is the reciprocal of the probability that the locations in this desired collection unit are in a subsampled assigned unit. Then let $w_{ij} = W_{ij}$ if desired collection unit *ij* is a sample unit after step 3 and $w_{ij} = 0$ otherwise; let c_{ij}, $j = 1, \ldots, N_i$, denote a set of constants satisfying

$$\sum_{j=1}^{N_i} c_{ij} = 1; \tag{3}$$

and let

$$w_i = \sum_{j=1}^{N_i} c_{ij} w_{ij} \ . \tag{4}$$

Then w_i clearly satisfies (1) since

$$E(w_{ij}) = 1, \quad j = 1, \ldots, N_i \, . \tag{5}$$

Furthermore, although (1) is satisfied for any set of $c_{ij}'s$ satisfying (3), for variance purposes we would like to reduce the variability of w_i, and consequently equalize the values of $c_{ij} W_{ij}$, $j = 1, \ldots, N_i$. Unless step 3, the subsampling step, is needed, this requires that $c_{ij} = p_{ij} / \sum_{k=1}^{N_i} p_{ik}$, where p_{ij} is the probability of selection of the assigned unit associated with the desired collection unit *ij*. Now, provided each of the assigned units associated with an actual collection unit are noncertainty units from the same sampling stratum, p_{ij} is proportional to the assigned employment, denoted A_{ij}, and hence

$$c_{ij} = A_{ij} / \sum_{k=1}^{N_i} A_{ik} \ . \tag{6}$$

Now A_{ij} is known for each sampled assigned unit *ij* that is associated with actual collection unit *i*. However, the assigned employment for nonsampled assigned units that are associated with this actual collection unit may not be known. Obtaining them would require extracting data from the UDB and, furthermore, there may be problems matching some assigned units to the UDB. Consequently instead of using (6) we let

$$c_{ij} = R_{ij} / R_i \ , \tag{7}$$

where R_i is the employment reported by the respondent for the *i*-th actual collection unit during data collection and R_{ij} is the reported employment for the *j*-th desired collection unit within the *i*-th actual reported unit. Note that R_{ij} need only be obtained for sampled desired collection units, since $c_{ij} w_{ij} = 0$, for all nonsampled desired collection units. Now, if

$$\sum_{k=1}^{N_i} R_{ik} = R_i \tag{8}$$

then, (7) l satisfies (3), and hence $\hat{Y}$ would be an unbiased estimator of *Y*. However (8) does not necessary hold. This is because if there are any birth locations that are reporting to the UI under numbers issued after the frame was constructed, they are not associated with any assigned unit. Since they are not part of the frame, and hence are not part of any desired collection unit, they do not contribute to $\sum_{k=1}^{N_i} R_{ik}$, but do contribute to R_i. Therefore, we should attempt to at least have the respondent exclude such locations from the actual collection unit. Note, however, if the data for birth locations is included in both $\hat{Y}_i$ and R_i then $E(w_i \hat{Y}_i) = Y_i$ provided the following both hold:

- The ratio $(\sum_k Y_{ik}) / (\sum_k R_{ik})$ where the summation is over birth locations in actual collection unit *i* is the same as this ratio where the summation is over non-birth locations in the actual collection unit.

- c_{ij} is computed using (7) and w_i using (4).

Consequently, if these assumptions hold for all actual collection units *i* that include birth units with new UDB unit numbers then (2) remains an unbiased estimator of *Y*.

Now, although R_i is always known for any responding actual collection unit, it is possible that the respondent will not be able to provide the values of R_{ij} for sample desired collection units, in which case in place of (7) we could use

$$c_{ij} = A_{ij} \; / \; R_i \; . \tag{9}$$

With this value of c_{ij}, (3) does not necessarily hold even if (8) does, since $A_{ij} \neq R_{ij}$ in general. However if the time lag is relatively short between the time of frame construction and the time of initial data collection, A_{ij} and R_{ij} generally do not differ by much provided desired collection unit *ij* consists of the same locations as the associated assigned unit. An example of a situation where the desired collection unit and the associated assigned unit may be different occurs when the assigned unit consists of several locations and some, but not all of them have been sold to a single owner. If desired collection unit *ij* consists of these sold locations and the new owner can neither provide data that does not also include their other locations, nor provide a separate value of R_{ij} for the bought locations, then substituting A_{ij} for the unknown R_{ij} will generally provide a poor approximation to R_{ij}. This is because A_{ij} includes the assigned employment of the locations of the original assigned units that were not sold in addition to those that were sold, while R_{ij} would include only the sold locations. In that case it may be best to treat the desired collection unit as a nonrespondent collection unit and apply the adjustment in step 5.

We have discussed several possible values for c_{ij}. The value of c_{ij} that we have been using in the NCS in the case when $w_{ij} > 0$ for all *j* is (6). In this case c_{ij} is known as the merge adjustment factor. Otherwise, we have been using (9). In this case c_{ij} is also known as an adjustment for "collected more than assigned" or a case of the "document adjustment factor," with the other case described in Step 5. In the case when $w_{ij} > 0$ for at least two *j*'s, the summation in (4) has also been described as the final weighting step for merges. The reason that we have not been using (7) is the difficulty in obtaining R_{ij} from a respondent who cannot separate out the data.

Until now we have been assuming that there is no nonresponse, either of entire assigned units or of desired collection units after splitting. Since there are both types of nonresponse, the weights W_{ij} actually used in the above formulas are the weights after Step 5 which include both types of nonresponse adjustments. As result (5) no longer holds in general and hence neither does (1) for the w_i defined by (4). Consequently, (2) is not an unbiased estimator of *Y*, but this an expected consequence of nonresponse adjustment.

In the remainder of the section we discuss birth samples, multiple panels and update interviews, that is interviews after the first or initiation interview. No final decisions have been made on any of these topics, so what follows is just our present thinking.

We next consider the birth sample. We are planning as we initiate a new panel, to simultaneously select a birth sample. The birth sample is a sample of UDB numbers which are on our new frame, but were not on the previous year's frame. A portion of the birth sample will be allocated to the new panel and these birth sample units can be treated as any other units in the new panel. The remainder of the birth sample will be allocated among the continuing panels. For each of these units it is important that the associated actual collection unit only include birth locations. There does not appear to be anyway to allow the actual collection units to include ongoing locations without creating bias problems in the estimates. This is because the birth units and the ongoing units are selected from separate frames.

In our panel design it is possible for the same actual collection unit to contribute to the estimates for more than one panel. The key rule we intend to follow for panel weighting is to remember to keep the weighting separate for each panel. For example, suppose a company has two physical locations each with different UI reporting numbers and will only provide data for the two locations combined. Suppose only location 1 is selected for panel 1 and only location 2 for panel 2. Then if, for example, (7) and (4) are used to weight the actual collection unit, the panel 1 weight for this company at initiation would be the assigned weight of location 1 for panel 1 times its share of the company's reported employment at the time of the initiation for panel 1. Similarly its panel 2 weight at initiation would be the assigned weight of location 2 for panel 2 times its share of the company's reported employment at the time of the initiation for panel 2.

Further complications arise during updates. If the actual collection unit is associated with multiple assigned units, the set of such assigned units can change over time due to births, deaths and change of ownership. Furthermore, if assigned units associated with an actual collection unit are selected into the sample from different panels, then the actual collection unit will be tabulated under more than one panel.

There are several ways of handling the weighting for problems encountered during updates. What appears to be the easiest, both conceptually and operationally, is to treat an actual collection unit from the previous year as if were an assigned sample unit for the current year's update. This means that the weight at the end of step 6 from the previous year applies to all physical locations which comprise an actual collection unit for that year. The weighting and following rules that applied at initiation then apply during update. For example, if an actual collection unit from the previous year subsequently buys a physical location that was part of a different actual collection unit that year and now reports for that additional physical location in addition to the previous year's set of set of locations, then this combined unit would have to be reweighted using (7) and (4) for example, with R_{ij} being the current reported employment. In this example there would be two j's, one corresponding to last year's locations and the other to the additional location. Note that the use of (6) instead of (7) in this computation may be impractical, because it would not be an easy task matching up current UDB numbers with the panel frames and the birth sample frames.

Note that this approach to weighting for updates has the following effect. All responding physical locations that are part of sampled assigned units will have a positive weight for the

initiation interview. This also holds true for physical locations that are part of an actual collection unit at initiation that contained a location that was part of a sample assigned unit. In each update all physical locations that had a positive weight for the previous interview will continue to have a positive weight, as will all locations that previously did not have a positive weight but are now part of an actual collection unit that consists of at least one location with a positive weight. Thus having a positive weight can be viewed as a spreading infection that a location catches by becoming part of an actual collection unit in which at least one location has the infection. Furthermore, once infected, there is no cure until the panel ends, other than through nonresponse or going out of business. Note also that any birth units, from birth samples selected and assigned to the panel in years prior to the update have a positive weight. These birth units are treated in the update weighting process like any units that were selected from the panel's original sampling frame.

3. Occupational Issues

Except in the case of COCs, which are described separately in the next section, once it has been determined for which actual collection units PSO sampling will be done through the six step process described in the previous section, the selection of defined occupations for the initiation interview themselves entail relatively few data collection issues that require sampling and weighting modifications. This is because, unlike the case for establishment sampling where the actual collection units can be quite different in certain situations from the original assigned units, the selection of employees from an employee list is relatively straightforward. There is, of course, some occupational nonresponse, which requires two stages of occupational nonresponse adjustment as described in Black, Ernst, and Tehonica (1997). In this section we discuss only one occupation selection issue for the initiation interview, the subsampling of departments for certain occupations. Then we discuss several occupational issues relating to the updates, where complications can occur. As is the case for establishment sampling, no final decisions have been made on how occupational issues relating to updates will be handled.

3.1 Subsampling of Departments

In Step 3 of Section 2 a procedure was presented for reducing respondent or interviewer burden by subsampling locations. We describe here another burden reducing procedure which is used at the occupational selection level. Ideally when an actual collection unit consists of multiple locations and/or departments, and an occupation is selected, data should be collected for all employees in the occupation in the collection unit. Occasionally, it is impossible to obtain data for an entire collection unit when an occupation is selected that is a dominant occupation in the unit. The respondent burden and collection burden can be overwhelming in some situations. For example, an elementary teacher in a school district or a nurse in a hospital could be quite burdensome in large school districts and hospitals. In these cases, methods of subsampling the occupation to particular departments or locations are used.

If the respondent is willing to give information by department or location , collection of the data occurs for the department or location in which the occupations that were selected reside.

For example, there are 10 schools within a school district and elementary teacher is sampled 4 times. Each selected elementary teacher resides in a different school. The collection for elementary teacher is limited to the 4 schools. No additional weighting adjustment is needed since the final employee weight or individual weight will take the number of employees collected for into account during its computation. That is, the reciprocal of the probability of selection of the defined occupation, which is a key component of this weight, is the PSO sampling interval divided by the number of employees in the defined occupation. If data is only collected for teachers in a specific school, for example, then that becomes the defined occupation from a sampling and weighting perspective.

3.2 Occupational Issues during Updates

There is an additional weighting complication that applies to updates that is not a concern for the initiation interview, namely, how to properly reflect the occupational selections in the weighting process when the actual collection units change composition over time. At initiation this is not a problem because the weighting approach given by (4) only is needed at the establishment level. That is, the contribution to an employee weight arising from the occupational selection simply reflects the reciprocal of the probability of selection of the defined occupation from the actual collection unit at initiation. This contribution after adjustment for occupational nonresponse is multiplied by the final establishment weight which would be obtained through (4).

To illustrate the complication for updates consider the following example. Suppose two physical locations are each selected as assigned units. At the time of an update the two locations have merged and the respondent is only willing to report data for the combined unit. If we use the occupational selections from each of the physical locations as the occupational selections for the combined unit, then any defined occupation that at the time of initiation was present in both samples would have two chances of selection. This can be reflected in the weighting by adopting the weighting approach in (4), with employee weights being used instead of occupational weights. That is, for each defined occupation w_{ij} would now be the corresponding employee weight for desired collection unit *ij* (a single location in this example) and w_i would be the employee weight for the defined occupation for actual reporting unit *i* (the two merged locations in this example).

Unlike the case for establishment weighting, (4) can lead to biased estimates even if, for example, (7) is used and (8) holds, since (5) need not hold when w_{ij} is an employee weight. For example, the defined occupation may not have been present at initiation in all desired collection units that comprise an actual collection unit, in which case $w_{ij} = 0$ always for the defined occupations for those desired collection units *ij* for which the defined occupation was not present at initiation. There does not appear to be any clear way around the bias problem in that situation. In addition, the defined occupation may be restricted to a department, division or physical location that is contained within one desired collection unit *ij*, in which case $w_{ik} = 0$ for $k \neq j$. In that case the weighting problem can be handled by letting $c_{ij} = 1$ and $c_{ik} = 0$ for $k \neq j$. Note, that this last situation appears unlikely to occur, since if the

respondent is able to supply data for a defined occupation that does not cross desired collection units then it is unlikely that we would have the problem of actual collection units that consist of more than one desired collection unit to begin with.

An alternative approach to handling occupations in the above merger example would be to reselect the defined occupations for the merged unit. In that case (4) need only be used at the establishment level and the resulting establishment weight w_i would be multiplied by the reciprocal of the probability of selecting the defined occupation during the reselection, together with the occupational nonresponse adjustments, to obtain the employee weight

Changes at the occupational level that require reweighting can occur for an update even when the actual collection unit at an update remains unchanged from the previous interview. For example, suppose a quote at initiation consists of all employees in a specific defined occupation in Department A of an establishment. Department A merges with Department B prior to the update and now data is collected from the merged department. If we do not adjust the weights to take into account the merger we will tend to overestimate employment. This situation can be handled by using (4) and (7), analogous to their use for establishment mergers. Here, however, R_{ij} will denote the employment in defined occupation i for the j-th department that is part of the merger; R_i is the employment in this occupation in the merged department; w_{ij} is the employee weight prior to the merger for the selection of defined occupation i from the j-th department and w_i is the employee weight after the merger for this occupation. As in the case of establishment mergers, w_{ij} is only nonzero for originally selected departments in the establishment for defined occupation i, ignoring here complications such as prior departmental mergers. Similarly, if the sampled department splits and data are collected only for a subsample of the original department, then this situation could be handled like subsampling at the establishment level as described in Step 3 of the establishment process.

4. COC Establishments

Another data collection problem in NCS is the collection of sampled units that belong to a large company which has a policy that requires collection of data for their establishments be done from a single respondent. This single respondent is normally located at the central office for the company. Because of this, these sampled units are referred to as central office collection establishments, or COC establishments. Conducting PSO separately at each sampled COC establishment of a company becomes burdensome to the respondent which can jeopardize cooperation from these types of companies. Because COC establishments are found in most of the NCS PSU samples, it is imperative that we get cooperation from these types of companies.

In the OCSP, the issues surrounding COC establishments in NCS were not found due to the use of a fixed job list. Each time a respondent was asked for information in OCSP, the same occupations were asked for and the burden became somewhat limited as they became more familiar with the OCSP fixed list and could provide data to BLS periodically for the same set

of jobs. In the NCS design, occupations are normally selected within each sampled establishment during PSO. Therefore, the single respondent for each COC needs to be contacted each time one of their establishments is sampled and collected for the NCS. In anticipation of the respondent cooperation issues with some sensitive COCs an alternate method of collection was needed. We took the OCSP fixed job list type of thinking into account when determining what alternate collection method could be used. We also knew that we were not able to obtain a single list of sampled establishments nationwide for each company since all of the PSU samples are not selected simultaneously. The inability to select all of the PSU samples at one time is due to workload constraints and the need to extract the most current information from the UDB to select the samples.

In consultation with regional office collection staff the following alternate method of collection for the COCs was proposed. Since each of the companies was willing to provide national employment counts on all occupations it was determined that PSO could be conducted on this national data for each company to create a fixed job list for its establishments. The size of the occupation sample is determined on a case by case basis taking into account the number of establishments nationwide, the employment count nationwide, and the number of occupations nationwide. This type of occupation selection is known as Central PSO (CPSO).

A major disadvantage of using CPSO is that when the fixed job list is used for each establishment, there is no guarantee that each establishment selected for NCS for the COC will include any of the occupations on the fixed list. This means that the estimated employment for a sample establishment may be 0 or may be many times larger than the PSO employment for the establishment. In NCS, when PSO is done separately at each sample establishment, the sampling and weighting methodology we use guarantees that the estimated employment for the establishment will always equal the total PSO employment, regardless of which occupations are selected. This is not the case when CPSO is conducted without the use of an extra weighting adjustment described below.

CPSO has been put to use in the NCS. A sample of occupations is selected pps without replacement using the national data for a COC. The data includes the nationwide employment for each occupation in the COC. Much of the national data for a company is split into divisions based on function or SIC. In these cases, a fixed job list for each division is selected. The establishments within these divisions are somewhat homogeneous in their occupational makeup. For example a retail company may have a few different types of entities such as retail stores, distribution centers, and a corporate office. Selecting an occupation sample for each of these divisions can decrease the negative impact this type of selection may bring.

Weighting of COC Establishments

As noted in the Introduction, the weight for each employee in a selected job in NCS is obtained by taking the product of the reciprocal of the probability of selecting the establishment, the reciprocal of the probability of selecting the job given that the establishment is selected, and nonresponse adjustment factors for establishment and occupation

nonresponse. For COCs there are some differences in these factors and an additional adjustment that is used only for COCs.

The reciprocal of the probability of selecting a COC is no different than for a non-COC establishment, since the COC procedure only impacts the occupation selections. However, during the nonresponse adjustment procedures, the establishments that used the alternate collection procedure, CPSO, are given an establishment nonresponse adjustment factor of 1 and also an occupation nonresponse adjustment factor of 1. They are not put into the nonresponse cells formed for the normal PSO schedules.

As for the reciprocal of the probability of selecting each defined occupation, the main point is that value is essentially independent of the set of employees at the individual establishment since it is determined by the selection from the CPSO list. For each certainty occupation, the value is 1. For each noncertainty occupation the value is the final PSO sampling interval divided by the number of employees on the nationwide COC list for the selected occupation. The final PSO sampling interval is the number of employees on the COC nationwide list in noncertainty jobs divided by the number of noncertainty jobs selected.

For each COC establishment a single employment adjustment factor (EAF) is calculated for the noncertainty sampled occupations collected in the CPSO schedules. This factor adjusts the nationwide based occupation component of the employee weight to account for what is found at an individual establishment so that the occupational component of the employee weights summed over all employees in a sampled job equals the establishment's PSO employment. For non-COC establishments the EAF is not needed since this equality always holds for such establishments without an EAF. For the certainty occupations, the EAF is 1.0000.

To obtain the EAF, first let PSOE denote the PSO employment for a COC establishment. Let n_C, n_S denote the number of certainty and selected noncertainty occupations, respectively, selected from the CPSO list. Let E_{Ci}, $i = 1,\ldots,n_C$, and E_{Si}, $i = 1,\ldots,n_S$, denote the employment in the establishment for the *i*-th certainty occupation and the *i*-th selected noncertainty occupation, respectively. Finally, let W_{Si}, $i = 1,\ldots,n_S$ denote the reciprocal of the probability of selection of the *i*-th selected noncertainty occupation during CPSO. The employment adjustment factor is then

$$EAF = \frac{PSOE - \sum_{i=1}^{n_C} E_i}{\sum_{i=1}^{n_S} \left(W_{Si} \times E_i\right)}$$

The numerator of this fractional factor is the establishment's PSO employment in noncertainty jobs. The denominator is the unadjusted estimate of the number of employees in the establishment in non certainty jobs. Note that if the establishment has no employment in any

of the selected noncertainty jobs, then the EAF is not defined since the denominator of the above expression is 0. Also, in some circumstances this factor can be very large which can result in a large increase in the variances of some estimates. We are therefore considering setting a maximum allowable value for this factor. Finally, like other adjustments of this type, this adjustment would introduce a bias into most estimates if they were not already biased.

5. Variance Estimation

In this section we discuss how the sampling and weighting issues discussed in the previous sections, such as multiple panels, actual collection units that differ from the desired collection units, splitting, subsampling, and COCs are taken into account in variance estimation. The variance estimation program we are using is based on the Taylor series linearization approach. The actual variance estimation formulas and other details of this program are detailed in Tehonica, Ernst, and Ponikowski (1997).

Multiple panels. We first linearize the estimator for which we wish to compute variance estimates. Since the samples in each panel are selected essentially independently, the variance of the linearized estimator is simply the sum of the variances of the contribution to the estimator from each panel. Note that this independence assumption is not quite true, because the birth samples are not selected independently for each panel, but ignoring this slight dependence in computing the variance estimates should have negligible consequences on their accuracy.

Multiple selections in a panel. As described in previous sections an actual collection unit may consists of more than one desired collection unit, and hence correspond to more than one assigned unit. In such situations, each sample assigned unit associated with an actual collection unit should be considered as a separate selection for variance purposes just as each selection of the same unit is considered a separate selection for variance purposes for simple random sampling with replacement. For example, if an actual collection unit i is selected three times, this means that $w_{ij} > 0$ for three of the desired collection units j that comprise this actual collection unit, where w_{ij} is as defined in Section 2 . Then the data for this actual collection unit would enter the variance estimation formula three times, each with a weight $c_{ij} w_{ij}$ corresponding to the respective desired collection unit. That is we use the weight before summing in (4).

Subsampling of establishments and splits. Step 3 of the six steps in Section 2, the subsampling of establishments step, and Step 4, the splits before PSO sampling step can impact the variance estimates. However, splits after PSO is basically a data collection issue, not a sampling issue, and does not impact the weighting or the variance estimates unless it leads to occupational nonresponse. The impacts of subsampling and splits depend on whether the original assigned unit is certainty or noncertainty as detailed below.

Subsampling and splits before PSO of noncertainty assigned units. All subsampling and splits arising from the same assigned unit combined together are considered a single unit for variance purposes. This is because in the variance estimation formula there is a term for each first stage unit. In the case of noncertainty establishments the first stage unit is the sample assigned unit and consequently neither the splitting or subsampling has an impact on the set of first stage units. Note, however, that the estimates that enter the variance formula for each first stage unit use the employee weights from all employees in selected occupations in any split and incorporate all stages of weighting including the components of the weights arising from the subsampling in step 3 and the component arising from the PSO selection which will vary with the split location.

Subsampling of certainty assigned units. Each subsampled unit which is noncertainty is treated in the variance estimation formula exactly like a noncertainty assigned unit ordinarily would be, that is there is a single separate term for each such unit. This is because the subsampled unit is the first stage unit in that case. Each subsample unit that is subsampled with certainty is treated as certainty assigned unit ordinarily would be, that is with each of the occupational selections arising from this subsampled unit considered as a first stage unit and represented by a separate term in the variance estimation formula.

Split of a certainty assigned unit before PSO sampling. Each split should be considered as a separate certainty unit in the variance estimation formula. This is because the set of occupational selections corresponding to each split is the first stage units for each split.

COCs. There is relatively little modification needed in the variance estimation formulas for COCs. Although the selection of occupations is different for COCs, the weights that the variance estimation program uses will incorporate this different method of selection. The biggest difference arises from the EAF. This too will be incorporated into the weights used in the variance estimation program. However, this factor is subject to sampling variability and incorporating its variability into the variance estimation requires an additional Taylor series linearization which is presented in Tehonica, Ernst and Ponikowski (1997).

References

Black, S. R., Ernst, L. R., and Tehonica, J. (1997), "Sample Design and Estimation for the National Compensation Survey," in *Proceedings of the Survey Research Methods Section,* American Statistical Association, to appear.

Cohen, S. H. (1997), "The National Compensation Survey: The New BLS Integrated *Proceedings of the Survey Research Methods Section,* American Statistical Association, to appear.

Ernst, L. R. (1989), "Weighting Issues for Longitudinal Household and Family Estimates," in *Panel Surveys*, eds. D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh, New York: John Wiley & Sons, pp. 139-159.

Tehonica, J., Ernst, L. R., and Ponikowski, C. H. (1997), "Summary of Estimation and Variance Specifications for the 1996 Albuquerque, NM COMP2000 Test Survey,"
Bureau of Labor Statistics memorandum to The Record, dated March 3.

Any opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics.