Appendix B. Survey Method and Reliability Statement for the May 2003 Occupational Employment Statistics Survey

The Occupational Employment Statistics (OES) survey is a mail survey measuring occupational employment and wage rates for wage and salary workers in nonfarm establishments in the 50 States and the District of Columbia. Guam, Puerto Rico, and the Virgin Islands are also surveyed but their data are not included in the national estimates.

About 6,500,000 establishments are stratified within State by substate area, industry, and employment size class. The substate areas include all officially defined metropolitan areas, and one or more balance areas are defined for each State (MSA/BOS areas). The North American Industry Classification System (NAICS) is used to stratify establishments by industry.

Probability sample panels of about 200,000 establishments are selected semiannually. Responses are obtained through mail and telephone contact. Respondents report the number of employees by occupation using the Standard Occupational Classification (SOC) system. For each occupation, the number of employees is distributed across 12 wage intervals.

Estimates are based on a rolling 6-panel (or 3-year) cycle. The total sample size when 6 panels are combined is approximately 1.2 million establishments. For the May 2003 survey about 79 percent of establishments responded, covering about 72 percent of weighted employment. National occupational employment (SOC) and wage rate estimates are made for all 3-digit NAICS codes, most 4-digit NAICS codes, and selected 5-digit NAICS codes. Subnational industry detail varies by state/MSA/BOS.

Occupational and Industrial Classification Systems

The occupational classification system. In 1999, the OES survey began using the U.S. Office of Management and Budget's (OMB's) new occupational classification system known as the Standard Occupational Classification (SOC) system. (See appendix A for a detailed description of the system.) This is the first occupational classification system that OMB has required for Federal statistical agencies. The survey uses the system to categorize workers across 22 major occupation groups in 1 of approximately 770 detailed occupations.

The industrial classification system. In 2002, the OES survey switched from the Standard Industrial Classification (SIC) system to the North American Industry Classification System (NAICS). More information about NAICS can be found on the BLS Web site at **www.bls.gov/bls/naics.htm** or in the *2002 North American Industry Classification System* manual. Each establishment is assigned a 6-digit NAICS code based on its primary activity.

Industrial scope and stratification. The survey covers the following NAICS industries:

11	Logging (1133), Support Activities for Crop Production (1152), and Support Activities for Animal Production (1152) *only*
21	Mining
22	Utilities
23	Construction
31-33	Manufacturing
42	Wholesale Trade
44-45	Retail Trade
48-49	Transportation and Warehousing
51	Information
52	Finance and Insurance
53	Real Estate and Rental and Leasing
54	Professional, Scientific, and Technical Services
55	Management of Companies and Enterprises
56	Administrative and Support, and Waste Management and Remediation Services
61	Educational Services
62	Health Care and Social Assistance
71	Arts, Entertainment, and Recreation
72	Accommodation and Food Services
81	Other Services (except Public Administration), *excluding* private households (814)
	Federal Government
	State Government
	Local Government

About 350 industry groups are used for stratification. Some are 5-digit NAICS "exceptions." Most are either entire 4-digit NAICS codes or the residual 4-digits with the 5-digit exceptions removed. "NAICS4/5" is a short term that is sometimes used to describe this particular grouping of industries.

Concepts

An ***establishment*** is generally a single physical location at which economic activity occurs (e.g., store, factory, farm, etc.). Each establishment is assigned a 6-digit NAICS code. When a single physical location encompasses two or more distinct economic activities, it is treated as separate establishments if separate payroll records are available and certain other criteria are met.

Employment is the number of workers who can be classified as full- and part-time employees, including workers on paid vacations or other types of leave; workers on unpaid short-term absences; salaried officers, executives, and staff members of incorporated firms; employees temporarily assigned to other units; and non-contract employees for whom the reporting unit is their permanent duty station regardless of whether that unit prepares their paychecks.

The OES survey includes all full- and part-time wage and salary workers in non-farm industries. Self-employed owners, partners in unincorporated firms, household workers, and unpaid family workers are excluded.

Occupations are classified based on work performed and on required skills. Employees are assigned to an occupation based on the work they perform and not on their education or training. For example, an employee trained as an engineer but working as a drafter is reported as a drafter. Employees who perform the duties of two or more occupations are reported in the occupation that requires the highest level of skill or in the occupation where the most time is spent if there is no measurable difference in skill requirements. ***Working supervisors*** (those spending 20 percent or more of their time doing work similar to the workers they supervise) are classified with the workers they supervise. ***Workers receiving on-the-job training, apprentices, and trainees*** are classified with the occupations for which they are being trained.

A ***wage*** is money that is paid or received for work or services performed in a specified period. Base rate pay, cost-of-living allowances, guaranteed pay, hazardous-duty pay, incentive pay such as commissions and production bonuses, tips, and on-call pay are included in a wage. Back pay, jury duty pay, overtime pay, severance pay, shift differentials, non-production bonuses, employer costs for supplementary benefits, and tuition reimbursements are excluded. Employers are asked to classify each of their workers into an SOC occupation and one of the following 12 wage intervals:

Interval	*Wages*	
	Hourly	*Annual*
Range A	Under $6.75	Under $14,040
Range B	$6.75 to $8.49	$14,040 to $17,679
Range C	$8.50 to $10.74	$17,680 to $22,359
Range D	$10.75 to $13.49	$22,360 to $28,079
Range E	$13.50 to $16.99	$28,080 to $35,359
Range F	$17.00 to $21.49	$35,360 to $44,719
Range G	$21.50 to $27.24	$44,720 to $56,679
Range H	$27.25 to $34.49	$56,680 to $71,759
Range I	$34.50 to $43.74	$71,760 to $90,999
Range J	$43.75 to $55.49	$91,000 to $115,439
Range K	$55.50 to $69.99	$115,440 to $145,599
Range L	$70.00 and over	$145,600 and over

Three-year Survey Cycle of Data Collection

The survey is based on a probability sample drawn from a universe of about 6,500,000 in-scope establishments stratified by geography, industry, and employment size. The sample is designed to represent all non-farm establishments in the United States.

Beginning with the November 2002 panel, the OES survey changed from an annual sample of 400,000 establishments to a semiannual sample of 200,000 establishments in May and November of each year in order to reduce seasonal bias. The semiannual samples are referred to as panels, and previous yearly samples are considered to be the equivalent of two panels. To the extent possible, privately owned units selected in any one panel will not be sampled again in the next five panels.

The survey is conducted over a rolling 6-panel (or 3-year) cycle. This is done in order to maintain adequate geographic, industrial, and occupational coverage. Over the course of a 6-panel (or 3-year) cycle, approximately 1.2 million establishments are sampled. For example, data collected in May 2003 are combined with data collected in November 2002, 2001, and 2000. For this transitional set of estimates, a subset of certainty units collected in 1999 was also used in the May 2003 estimates. The May 2003 and November 2002 data are semiannual samples while the 2001 and 2000 data are annual samples—the equivalent of 6 panels when combined. Data from 1999 are added to provide complete coverage of strata with the largest establishments (250+ employees based on maximum size).

For a given panel, survey questionnaires/schedules are initially mailed out to almost all sampled establishments. State Employment Security Agency (SESA) staff may make "early" personal visits to some of the larger establishments. Two additional mailings are sent to nonrespondents at approximately 3-week intervals. Telephone or personal visit follow-ups are made to nonrespondents critical to the survey because of their size.

A census is obtained semiannually, representing May and November employment, of Federal Government establishments (annually prior to the November 2002 panel). Data for Federal Government employment and wages are collected at the end of the data collection process from the U.S. Office of Personnel Management. A semiannual census of workers is also obtained from the United States Postal Service (USPS). An annual census, representing November data, is obtained for State government units. These census reports are broken out in fine geographic detail. The Federal and State

census reports only have default industry codes that indicate "Federal government" or "State government." Local government units are subject to probability sampling, but the reports only have a default industry code indicating "local government."

Sampling Procedures

The frame

The sampling frame, or universe, is a list of about 6,500,000 in-scope non-farm establishments that file unemployment insurance (UI) reports to the State Employment Security Agencies. Virtually all establishments are required to file these reports with the notable exception of establishments in Guam and rail transportation (NAICS 4821). Every quarter a sampling frame list is created by combining all the State lists into a single file called the Longitudinal Data Base (LDB). For the 1999 sample, the sampling frame was the 1998/2nd quarter LDB file; for the 2000 sample, it was the 1999/2nd quarter LDB file; for the 2001 sample, it was the 2000/4th quarter LDB file; for the November 2002 sample, it was the 2001/4th quarter LDB file; and for the May 2003 sample, it was the 2002/2nd quarter LDB file. The LDB files are also supplemented with a frame covering Guam and rail transportation (NAICS 4821).

Stratification

The frame is stratified geography-by-industry-by-size.

- The geographic stratification used is MSA/BOS within state. All officially defined metropolitan areas are used, and each State is allowed to define 1-6 balance-of-State areas.
- The industry stratification is the approximately 350 NAICS4/5 industry groups.
- Stratification uses seven employment size-class (SC) ranges: 1-4, 5-9, 10-19, 20-49, 50-99, 100-249, and 250+. The size of an establishment is based on its maximum monthly employment taken from the most recently available 12 months of administrative (universe) data.

At any given time there are about 550,000 nonempty MSA/BOS-by-NAICS4/5-by-SC strata on the frame. When comparing nonempty strata between frames, there are substantial frame-to frame differences. The differences are primarily due to the normal birth/death process and normal establishment growth/shrinkage. Some differences are due to NAICS reclassification and changes in geographic location assigned to establishments.

Certainty and virtual certainty units

Federal Government and USPS units are certainty units since a census is obtained for every panel. For State government units a census is obtained every other panel (representing November employment). Technically, the State units are not "certainty" units since data are not obtained for every panel; the term "virtual certainty" is used. The term "virtual certainty" also applies to the very largest units in the 250+ size class. All of the largest units are included once in the 6-panel survey cycle, if possible. (Sometimes reinterviewing a few of these largest units must be delayed as a result of budget considerations.)

Allocation of the sample to strata

For each state, a sample panel of establishments is selected within the MSA/BOS-by-NAICS4/5-by-SC stratification. Within a state, the sample is allocated in a manner that equalizes the expected relative standard error of typical occupational employment in each MSA/BOS-by-NAICS4/5 cell. Within each cell, the sample is allocated across the size classes in a manner that minimizes the variance of the average typical occupational employment estimate.

Sampling using PRNs

Permanent random numbers (PRNs) are used in the sample selection process. Each establishment in the sampling frame is assigned a PRN. The reason for using PRNs in sampling is that it gives us an easy method to limit sample overlap between the OES survey and other large surveys conducted by the Bureau of Labor Statistics.

Sample selection using PRNs can be implemented in several ways. For OES, a specific PRN value is designated as a "start" point in a stratum. Beginning with this "start" point, *n* establishments in the stratum are sequentially selected into the sample where *n* denotes the number of establishments to be sampled.

Panel weights (sampling weights)

Sampling weights are assigned so that each panel, when sampled establishments are weighted, will roughly represent the entire universe of establishments.

Federal Government, USPS, and State government units are assigned panel weights of 1. Other sampled establishments, including virtual certainties, are assigned design-based panel weights. For a stratum with n establishments sampled from N frame establishments, weight N/n is assigned to each of the n sampled establishments. N/n is the inverse of the panel probability of selection within the panel.

National sample counts

The combined sample for the May 2003 survey is considered to be the equivalent of a combined 6-panel sample. Approximately 1/6 of the combined sample comes from each of the semiannual sample panels for May 2003 and November 2002. Approximately 2/6 of the combined sample comes from the 2001 sample (a 2-panel equivalent) and another 2/6 from the 2000 sample (also a 2-panel equivalent).

Sample allocation resulted in initial sample sizes of:
199,587 establishments for May 2003
201,016 establishments for November 2002
405,655 establishments for 2001 (2-panel equivalent)

406,876 establishments for 2000 (2-panel equivalent)

In addition, 3,616 certainty units from 1999 were added to the sample to provide complete coverage of the certainty strata. The *combined* initial sample size for the May 2003 estimates is approximately 1,200,981 establishments. The combined count avoids double/triple-counting by appropriately subtracting out Federal and State Government establishments. For Federal Government establishments only the May 2003 census is counted (subtract out November 2002, 2001, and 2000). For State government establishments only the November 2002 census is counted (subtract out 2001 and 2000; no census of State government in May 2003).

Response and Nonresponse

Response

Of the 1,200,981 establishments in the combined initial sample, 1,099,307 were viable establishments. That is, they were not out-of-scope or out-of-business. Of the viable establishments, 863,182 responded and 236,120 were classified as nonrespondents. The establishment response rate is 78.5% (863,182/1,099,307). The response rate in terms of weighted sample employment is 72.0%.

Nonresponse

Nonresponding establishments are accounted for in the OES survey by a two-step imputation process.

- *Step 1, Occupational employment staffing pattern:* For each nonrespondent, a staffing pattern is imputed using a nearest-neighbor "hot deck" imputation method. The procedure links a donor responding establishment to each nonrespondent. For example, for the May 2003 survey, possible donors were the respondents from the May 2003, November 2002, and 2001 samples. The nearest-neighbor hot deck procedure for OES searches within defined cells for the donor that most closely resembles a nonrespondent in terms of geographic area, industry, and employment size. At first, a donor with approximately the same employment size is sought within the same MSA/BOS and 5-digit NAICS as the nonrespondent. The area/industry parameters of the donor pool are successively widened until a suitable donor is found. Limits are placed on the number of times a donor can be used. For a nonrespondent, its donor is used to impute (simulate) a response for the occupational employment data or staffing pattern. The donor's staffing pattern distribution is used for the nonrespondent but the level is adjusted to be appropriate for the nonrespondent's known employment size.

- *Step 2, Wage distribution:* A variation of mean imputation is used to simulate a wage distribution for each nonrespondent. Imputation cells are defined by geographic area, industry, and size class. Responding establishments in each cell are used to compute, for each occupation, a distribution across the 12 wage intervals. For nonrespondents in the cell, those wage distributions are applied to already imputed occupational employment. If a cell has insufficient response to compute a distribution, the cell is expanded into adjacent areas, industries, or size classes until sufficient response is achieved.

Occasionally a responding establishment reports occupational employment but not a distribution across the wage intervals for all or some occupations. In this situation, the imputation procedure described in step 2 is used to impute an occupational wage distribution.

Combining and benchmarking data for occupational employment estimates

Reweighting for the combined sample

Employment and wage rate estimates are computed using a rolling 6-panel (3-year) sample. For example, estimates are made using data from the May 2003, November 2002, 2001, and 2000 samples plus a small number of large virtual certainties held over from the 1999 sample. Establishments in each sample are weighted independently to represent the universe at the time it was selected. When panels are combined, each sampled establishment is reweighted so that the aggregate sample represents the universe.

Only the most recent Federal Government, USPS, and State government censuses are retained (with a certainty weight set to 1). The weight of all large virtual certainties is set to 1.

Noncertainties are analyzed stratum-by-stratum. The original single-panel sampling weights are set so that responses from a stratum can be weighted up to represent the entire stratum. In the simplest case, 6 panels are combined and all 6 have sample units for a particular stratum. Since a simple summation of single-panel weights would represent the stratum 6 times, the combined sample weight of each establishment is set equal to its single-panel sampling weight divided by 6. It is most common for some panels to have no sample. For example, if only 2 of 6 panels have sample for a stratum, then the single-panel sampling weights are divided by 2.

Benchmarking to QCEW employment

A ratio estimator is used to develop estimates of occupational employment. The auxiliary variable used is the average of the most recent May and November employment totals extracted from BLS' Quarterly Census of Employment and Wages (QCEW)—May 2003 and November 2002 for estimates made from the combined sample for the May 2003 survey. In order to balance the States' need for estimates at differing levels of geographic and industrial aggregation, the ratio adjustment process is carried out through a series of four hierarchical ratio adjustments. The procedure is com-

monly called benchmarking and the ratio adjustments are called benchmark factors (BMFs).

The first of the four hierarchical benchmark factors is calculated within states for cells defined MSA/BOS by NAICS4/5 by employment size class (4 size classes). If any first level BMF is out of range, it is reset to a predetermined maximum or minimum value. First-level BMFs are calculated in the following manner:

h = MSA/BOS by NAICS4/5
H = State by NAICS4/5
s = employment size classes (1-19, 20-49, 50-249, or 250+)
S = 1 of 2 aggregate employment size classes (1-49, 50+)
M = average of May and November QCEW
w_i = combined sample weight for establishment i
x_i = total establishment employment
$\text{BMF}_{\min}$ = a parameter, the lowest value allowed for BMF
$\text{BMF}_{\max}$ = a parameter, the highest value allowed for BMF

$$\beta_{hs} = \left(M_{hs} \Big/ \sum_{i \in hs} w_i x_i \right), \quad \beta_{hS} = \left(M_{hS} \Big/ \sum_{i \in hS} w_i x_i \right), \quad \beta_h = \left(M_h \Big/ \sum_{i \in h} w_i x_i \right),$$

then

$$BMF_{1,\,hs} = \begin{cases} \beta_{hs}, \text{ if all } \beta_{hs} \text{ within } h \text{ are bounded by } (\text{BMF}_{\min}, \text{BMF}_{\max}), \\ \beta_{hS}, \text{ if all } \beta_{hS} \text{ within } h \text{ are bounded by } (\text{BMF}_{\min}, \text{BMF}_{\max}), \\ \text{BMF}_{\min}, \text{ if } \beta_h < \text{BMF}_{\min}, \\ \text{BMF}_{\max}, \text{ if } \beta_h > \text{BMF}_{\max}, \\ \beta_h \text{ otherwise} \end{cases}$$

Second-level BMFs are calculated at the State by 4-digit NAICS cell level by summing the product of the combined sample weight and the first level BMF for each establishment in the cell. Second level BMFs account for the portion of universe employment that is not adequately represented by weighted employment after first-level benchmarking. In particular, some universe MSA/BOS by NAICS4/5 by size class cells have no sample and are not adequately represented by the weighted sample after first-stage benchmarking. Trimming first-level BMFs also causes over/under coverage that needs second-level benchmarking. Second-stage benchmarks are calculated as follows:

$$\beta_H = \left(M_H \Big/ \sum_{hs \in H} \sum_{i \in hs} w_i x_i BMF_{1,\,hs} \right)$$

$$BMF_{2,\,H} = \begin{cases} \text{BMF}_{\min}, \text{ if } \beta_H < \text{BMF}_{\min}, \\ \text{BMF}_{\max}, \text{ if } \beta_H > \text{BMF}_{\max}, \\ \beta_H \text{ otherwise} \end{cases}$$

Third- and fourth-level BMFs are calculated in a similar manner. The third-level BMF calculation of $\text{BMF}_{3,H}$ uses combined sample weights adjusted through second-level benchmarking. The fourth-level BMF calculation of $\text{BMF}_{4,H}$ uses combined sample weights adjusted through third-level benchmarking.

A final benchmark factor for each establishment, BMF_i, is calculated as the product of the four hierarchical ratio adjustment factors. That is, $\text{BMF}_i = \text{BMF}_1 * \text{BMF}_2 * \text{BMF}_3 * \text{BMF}_4$. A final weight value is calculated as the product of the combined sample weight and the final benchmark factor.

Occupational employment estimates

The final weights are used to calculate estimates of occupational employment that are benchmarked to QCEW employment. The May 2003 survey, for example, is benchmarked to the average of May 2003 and November 2002 QCEW. Estimates for a cell are produced simply by summing up the desired reported data for each establishment multiplied by the establishment's final weight.

The equation below is used to calculate occupational employment estimates at the MSA/4-digit NAICS cell level.

$$\hat{X}_{ho} = \sum_{i \in h} \left(w_i \; BMF_i \; x_{io} \right)$$

o = occupation
h = reported 4-digit NAICS code in an MSA
w_i = adjusted sample weight for establishment i
BMF_i = final benchmark factor applied to establishment i
x_{io} = reported employment for occupation o in establishment i
$\hat{X}_{ho}$ = estimated employment for occupation o in the MSA/ 4-digit NAICS cell

The estimated employment for an occupation at the MSA/all-industry level can be obtained by summing the occupational employment estimate $\hat{X}_{ho}$ across all the 4-digit NAICS industries in the MSA.

$$\hat{X}_o = \sum_{h=1}^{L_h} \hat{X}_{ho}$$

Lh = # of 4-digit NAICS reporting occupation o in the MSA

However, the estimate can be made directly simply by summing up the data for the appropriate establishments multiplied by their final weights.

Wage rate estimation

Externally derived factors are used in wage rate estimation:

- Mean wage rates for each of the 12 wage intervals
- Wage updating or aging factors

Occupational wage data reported in the OES are grouped data. Individual rage rates are not collected for the workers. Instead, we obtain the number of workers in an occupation who are paid wages within each of 12 wage intervals. For example, an establishment might report that it employs 10 secretaries: 2 in wage interval B, paid wages between $6.75 and $8.49 per hour; 6 in wage interval D, paid wages between $10.75 and $13.49 per hour; and 2 in wage interval E, paid wages between $13.50 and $16.99 per hour. Simple arithmetic mean formulas cannot be used to get valid estimates of means when data are grouped. For valid estimates of means, standard formulas for grouped data need an approximately unbiased average value within each group.

Data from several sample panels with different reference dates are used to produce OES wage estimates. Sample panels have different reference periods and the wage data are not equivalent in real-dollar terms. Data collected prior to the current survey reference period need to be updated or aged to approximate the latest reference period. For the May 2003 survey, for example, wage data from November 2002, 2001, 2000, and 1999 samples need to be aged.

Determining a mean wage rate for each interval

The average hourly wage rate for all workers in any given wage interval cannot be derived from collected OES data. It is approximated externally using data from the BLS National Compensation Survey (NCS). The mean hourly wage rate for interval L, the upper, open-ended interval, is calculated after excluding wage data for pilots, an occupation that accounts for a large proportion of NCS employment in interval L. Because pilots work much fewer hours than other occupations, their hourly wage rates are naturally much higher than other occupations. The mean hourly wage rate for interval L, without pilots, is calculated separately for each survey reference period then averaged.

Wage aging process

Aging factors are developed from BLS' Employment Cost Index (ECI) survey. The ECI survey measures the rate of change in compensation from a past survey reference period (4th quarter 2000, for example) to the current survey reference period (2nd quarter 2003, for example) for nine major occupational groups.

Mean hourly wage rate estimates

Mean hourly wage is the total hourly wages for an occupation divided by its weighted survey employment. Estimates of mean hourly wage are calculated using a standard grouped data formula that is modified to utilize ECI aging factors.

$$\hat{R}_o = \frac{\sum_{z=t-4}^{t} \left(\sum_{i \in z} w_i \; BMF_i \; \hat{y}_{i\,o} \right)}{\hat{X}_o}$$

$$\hat{y}_{i\,o} = u_{zo} \sum_r x_{ior} c_{zr} \qquad (i \in z)$$

o = occupation

$\hat{R}_o$ = mean hourly wage rate for occupation o

z = year (or panel)

t = current panel

w_i = combined sampling weight for establishment i

$\hat{y}_{i\,o}$ = unweighted total hourly wage estimate for occupation o in establishment i

r = wage interval

$\hat{X}_o$ = estimated employment for occupation o

x_{ior} = reported employment for occupation o in establishment i in wage interval r (note that establishment i reports data for only one panel z or one year z)

u_{zo} = ECI aging factor for year (or panel) z and occupation o

c_{zr} = mean hourly wage, interval r panel z (or year z)

In this formula, c_{zr} represents the mean hourly wage of interval r in panel (or year) z. The mean is determined externally using data from the Bureau's NCS survey. Research is conducted at periodic intervals to verify the continued utility of this updating procedure.

Percentile hourly wage rate estimates

The p-th percentile hourly wage rate for an occupation is the wage where p percent of all workers earn that amount or less *and* where (100-p) percent of all workers earn that amount or more. The wage interval containing the p-th percentile hourly wage rate is located using a cumulative frequency count of employment across all wage intervals. After the targeted wage interval is identified, the p-th percentile wage rate is then estimated using a linear interpolation procedure.

$$pR_o = L_r + \frac{j}{f_r}(U_r - L_r)$$

pR_o = p-th percentile hourly wage rate for occupation o

r = wage interval that encompasses pR_o

L_r = lower bound of wage interval r

U_r = upper bound of wage interval r

f_r = number of workers in interval r

j = difference between the number of workers needed to reach the p-th percentile wage rate and the number of workers needed to reach the L_r wage rate

Annual wage rate estimates

These estimates are calculated by multiplying hourly wage rate estimates (mean or p-th percentile) with a "year-round, full time" figure of 2,080 hours (52 weeks x 40 hours) per year. The estimates, however, may not represent mean annual pay if the workers work more or less than 2,080 hours per year.

Alternatively, some workers are paid based on an annual amount but do not work the usual 2,080 hours per year. Since the survey does not collect the actual number of hours worked, hourly wage rates cannot be derived with any reasonable degree of confidence from the annual rates.

Confidentiality

BLS has a strict confidentiality policy that ensures that the survey sample composition, lists of reporters, and names of respondents will be kept confidential. Additionally, the policy assures respondents that published figures will not reveal the identity of any specific respondent and will not allow the data of any specific respondent to be imputed. Each published estimate is screened to ensure that it meets these confidentiality requirements. The specific screening criteria are not listed in this publication to further protect the confidentiality of the data.

Variance estimation

Occupational employment variance estimation

A subsample replication technique called the "jackknife random group" is used to estimate variances of occupational employment. In this technique, each sampled establishment is assigned to one of G random groups. Using the data in these groups, G subsamples are formed from the parent sample. Each subsample is reweighted to represent the entire universe.

For an occupational employment total, G estimates of total occupational employment ($\hat{X}_{hjog}$) are calculated, one employment estimate per subsample. Then the variability among these G estimates is calculated to obtain an estimate of variance. For example, an occupational employment variance estimate for 4-digit NAICS *h and* reported size class *j* is calculated as follows.

$$v(\hat{X}_{hjo}) = \frac{\sum_{g=1}^{G} (\hat{X}_{hjog} - \bar{\hat{X}}_{hjo})^2}{G(G-1)}$$

$v(\hat{X}_{hjo})$ = estimated variance of $\hat{X}_{hjo}$

G = number of random groups

$\hat{X}_{hjo}$ = estimated employment of occupation *o* in NAICS *h* and size class *j*

$\hat{X}_{hjog}$ = estimated employment of occupation *o* in NAICS *h*, size class *j*, and subsample *g*

$\bar{\hat{X}}_{hjo}$ = estimated mean employment for occupation *o* in NAICS *h* and size class *j* based on the G subsamples (Note: a finite population correction factor is applied to the terms $\hat{X}_{hjog}$ and $\bar{\hat{X}}_{hjo}$.)

The variance for an occupational employment estimate at the reported 4-digit NAICS *h* level is obtained by summing the variances $v(\hat{X}_{hjo})$ across all reported size classes *j* in NAICS *h*.

$$v(\hat{X}_{ho}) = \sum_{j \in h} v(\hat{X}_{hjo})$$

Similarly, the variance for an occupational employment estimate at the reported 3-digit NAICS level *H* is obtained by summing the variances $v(\hat{X}_{ho})$ across all reported 4-digit NAICS *h*'s within the 3-digit NAICS.

$$v(\hat{X}_{Ho}) = \sum_{h \in H} v(\hat{X}_{ho})$$

Occupational mean wage variance estimates

Because the OES wage data are collected in intervals (grouped), we do not capture the exact wage of each worker. Therefore, some components of the wage variance are approximated using factors developed from NCS data. A *Taylor Linearization* technique is used to develop a variance estimator appropriate for OES mean wage estimates. The primary component of the mean wage variance, which accounts for the variability of the observed sample data, is estimated using the standard estimator of variance for a ratio estimate. This component is the first term in the formula given below:

$$v(\hat{R}_o) = \left(\begin{array}{l} \frac{1}{\hat{X}_o^2} \left(\sum_h \left\{ \frac{n_{ho}(1-f_{ho})}{n_{ho}-1} \right\} \left\{ \sum_{i \in h} w_i^2 (q_{io} - \bar{q}_{ho})^2 \right\} \right) + \\ \sum_r \theta_{or}^2 \sigma_{cr}^2 + \frac{1}{\hat{X}_o^2} \sum_r \left(\sum_{i=1}^{n_o} (w_i x_{ior})^2 \right) \sigma_{er}^2 + \frac{1}{\hat{X}_o} \sum_r \theta_{or} \sigma_{\omega r}^2 \end{array} \right)$$

$\hat{R}_o$ = estimated mean wage for occupation *o*

$v(\hat{R}_o)$ = estimated variance of $\hat{R}_o$

$\hat{X}_o$ = estimated occupational employment for occupation *o*

h = stratum (area/industry/size class)

f_{ho} = sampling fraction for occupation *o* in stratum *h*

n_{ho} = number of sampled establishments that reported occupation *o* in stratum *h*

w_i = sampling weight for establishment *i*

q_{io} = $\left(\hat{y}_{io} - \hat{R}_o x_{io}\right)$ for occupation *o* in establishment i

$\hat{y}_{io}$ = estimated total occupational wage in establishment *i* for occupation *o*

x_{io} = reported employment in establishment *i* for occupation *o*

$\overline{q}_{ho}$ = mean of the q_{io} quantities for occupation *o* in stratum *h*

θ_{or} = proportion of employment within interval *r* for occupation *o*;

x_{ior} = reported employment in establishment *i* within wage interval *r* for occupation *o*

$\left(\sigma_{cr}^2, \sigma_{er}^2, and\ \sigma_{\omega r}^2\right)$ Within wage interval *r,* these are estimated using the NCS and respectively represent the variability of the wage value imputed to each worker; the variability of wages across establishments; and the variability of wages within establishments.

Reliability of the estimates

Estimates developed from a sample may differ from the results of a census. An estimate based on a sample survey has two types of error—sampling error and nonsampling error. (A census would have only nonsampling error, but the nonsampling errors for a census and a survey designed to make the same estimates can be very different.)

Nonsampling error

This type of error is attributable to several causes, such as errors in the sampling frame; an inability to obtain information for all establishments in the sample; differences in respondents' interpretation of survey question; an inability or unwillingness of the respondents to provide correct information; errors made in recording, coding, or processing the data; and errors made in imputing values for missing data. Explicit measures of the effects of nonsampling error are not available.

Sampling errors

When a sample, rather than an entire population, is surveyed, estimates differ from the true population values that they represent. This difference, or sampling error, occurs by chance and its variability is measured by the variance of the estimate or the standard error of the estimate (square root of the variance). The relative standard error is the ratio of the standard error to the estimate itself, and is often called a coefficient of variation, especially when it is expressed as a percent of the estimate.

Estimates of sampling errors for occupational employment and mean wage estimates are provided in this publication to allow data users to determine if estimates are reliable enough for their needs. Only a probability-based sample can be used to calculate estimates of sampling error from the sample itself. The formulas used to estimate OES variances are adaptations of formulas appropriate for the survey design used.

The particular sample used in this survey is one of a large number of many possible samples of the same size that could have been selected using the same sample design. Sample estimates from a given design are said to be unbiased when an average of the estimates from all possible samples would yield, hypothetically, the true population value. In this case, the sample estimate and its standard error can be used to construct approximate confidence intervals, or ranges of values that include the true population value with known probabilities. To illustrate, if the process of selecting a sample from the population was repeated many times, if each sample was surveyed under essentially the same unbiased conditions, and an estimate of its standard error made from each sample, then:

1. Approximately 68 percent of the intervals from one standard error below to one standard error above the estimate would include the true population value. This interval is called a 68-percent confidence interval.

2. Approximately 90 percent of the intervals from 1.6 standard errors below to 1.6 standard errors above the estimate would include the true population value. This interval is called a 90-percent confidence interval.

3. Approximately 95 percent of the intervals from 2 standard errors below to 2 standard errors above the estimate would include the true population value. This interval is called the 95-percent confidence interval.

4. Almost all (99.7 percent) of the intervals from 3 standard errors below to 3 standard errors above the estimate would include the true population value.

For example, suppose that an estimated occupational employment total is 5,000, with an associated estimate of relative standard error of 2.0 percent. Based on these data, the standard error of the estimate is 100 (2 percent of 5,000). To construct a 95-percent confidence interval, add and subtract 200 (twice the standard error) from the estimate: (4,800, 5,200). Approximately 95 percent of the intervals constructed in this manner will include the true occupational employment if survey methods are nearly unbiased.

Estimated standard errors should be taken to indicate the magnitude of sampling error only. They are not intended to measure nonsampling error, including any biases in the data.

Particular care should be exercised in the interpretation of small estimates or of small differences between estimates when the sampling error is relatively large or the magnitude of the bias is unknown.

Quality control measures

Several edit and quality control procedures are used to reduce nonsampling error. For example, completed survey questionnaires are checked for data consistency. Follow-up mailings and phone calls are sent out to nonresponding establishments to improve the survey response rate. Response analysis studies are conducted to assess the respondents' comprehension of the questionnaire. (See the section below for additional information on the quality control procedures used by the OES survey.)

The OES survey is a Federal-State cooperative effort that enables States to conduct their own surveys. A major concern with a cooperative program such as OES is to accommodate the needs of BLS and other Federal agencies, as well as State-specific publication needs, with limited resources while simultaneously standardizing survey procedures across all 50 States, the District of Columbia, and the U.S. territories. Controlling sources of nonsampling error in this decentralized environment can be difficult. One important computerized quality control tool used by the OES survey is the Survey Processing and Management (SPAM) system. It was developed to provide a consistent and automated framework for survey processing and to reduce the workload for analysts at the State, regional, and national levels.

To ensure standardized sampling methods in all areas, the sample is drawn in the national office. Standardizing data processing activities such as validating the sampling frame, allocating and selecting the sample, refining mailing addresses, addressing envelopes and mailers, editing and updating questionnaires, conducting electronic review, producing management reports, and calculating employment estimates have resulted in the overall standardization of the OES survey methodology. This has reduced the number of errors on the data files and the time needed to review them.

Other quality control measures used in the OES survey include:

- Follow-up solicitations of nonrespondents, especially critical or large nonrespondents;
- Review of schedules to verify the accuracy and reasonableness of the reported data;
- Adjustments for atypical reporting units on the data file;
- Validation of the benchmark employment figures and of the benchmark factors; and
- Validation of the analytical tables of estimates at the NAICS4/5 level.