Appendix B. Survey Methods and Reliability of the May 2004 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 also are surveyed, but their data are not included in national estimates.

About 6.5 million establishments are stratified within State by substate area, industry, and employment size class. Substate areas include all officially defined metropolitan areas and one or more residual balance-of-State areas (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 their number of employees by occupation across 12 wage bands. The Standard Occupational Classification system (SOC) is used to define occupations.

Estimates of occupational employment and occupational wage rates are based on a rolling six-panel (or 3-year) cycle. The total sample size when six panels are combined is approximately 1.2 million establishments.

Occupational and industrial classification systems

The occupational classification system

The U.S. Office of Management and Budget's Standard Occupational Classification system (SOC) is used to define occupations. The survey uses the system to categorize workers across 22 major occupation groups spanning approximately 801 detailed occupations. See appendix A for a detailed description of the system.

The industrial classification system

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

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

11	Logging (1133), support activities for crop production (1151), 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	Healthcare and social assistance
71	Arts, entertainment, and recreation
72	Accommodation and food services
81	Other Services, except Public Administration [private households (814) are excluded]
	Federal Government (OES-designated code 999100)
	State Government (OES-designated code 999200)
	Local Government (OES-designated code 999300)

These sectors are stratified into approximately 350 industry groups. Most groups are entire four-digit NAICS codes. The rest are either *stand-alone* five-digit NAICS codes or residual four-digit NAICS codes with the *stand-alone* five-digit codes removed. "NAICS4/5" is shorthand that is used to describe this particular grouping of industries.

Concepts

An ***establishment*** is generally a single physical location at which economic activity occurs (for example, store, factory, farm, etc.). Each establishment is assigned a six-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 refers to the number of workers who can be classified as full- or 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 noncontract 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 nonfarm 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 that of the workers they supervise) are classified with the workers they supervise. ***Workers receiving on-the-job training and apprentices*** 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. Included in a wage are base-rate pay; cost-of-living allowances; guaranteed pay; hazardous-duty pay; and incentive pay, such as commissions and production bonuses, tips, and on-call pay. Excluded are back pay, jury duty pay, overtime pay, severance pay, shift differentials, nonproduction bonuses, employer costs for supplementary benefits, and tuition reimbursement. Employers are asked to classify each of their workers into an SOC occupation and 1 of the following 12 wage intervals:

	Wages	
Interval	***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

3-year survey cycle of data collection

The survey is based on a probability sample drawn from a universe of about 6.5 million in-scope establishments stratified by geography, industry, and employment size. The sample is designed to represent all nonfarm 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. Semiannual samples are referred to as panels. Previous yearly samples are 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 succeeding panels.

The survey is conducted over a rolling six-panel (or 3-year) cycle. This is done to provide adequate geographic, industrial, and occupational coverage. Over the course of a six-panel (or 3-year) cycle, approximately 1.2 million establishments are sampled. In this cycle, data collected in May 2004 are combined with data collected in November 2003, May 2003, November 2002, and 2001 (the equivalent of two panels). In the transition from annual to semiannual sampling, some large establishments from the 2000 sample were added to provide complete coverage of establishments with 250 or more workers (based on maximum employment).

For a given panel, survey questionnaires/schedules are initially mailed out to almost all sampled establishments. State workforce agency staff may make personal visits to some of the larger establishments. Two or three additional mailings are sent to nonrespondents at approximately 3-week intervals. Telephone or personal visit followups are made to nonrespondents considered critical to the survey because of their size.

Periodic censuses are taken of Federal and State Government.

- A census of Federal Government and the U.S. Postal Service (USPS) is conducted semiannually in June and December. Prior to November 2002, a census of Federal Government and the USPS was taken annually. Employment and wage data for these industries are collected from the U.S. Office of Personnel Management (OPM). Data from only the most recent panel is retained for use in OES estimates.

- A census of State government is conducted annually in November. State-owned establishments are included, except for schools and hospitals. In most States, a consolidated report of employment and wage data is obtained for each MSA/BOS level.

- A probability sample, not a census, is taken of local government units. Local-government-owned establishments are included, excluding schools and hospitals.

Schools and hospitals owned by State or local government are sampled along with their counterparts in the private sector.

Sampling Procedures

The frame

The sampling frame, or universe, is a list of about 6.5 million in-scope nonfarm establishments that file unemployment insurance (UI) reports with the State workforce agencies. Virtually all establishments are required to file these reports, with the exception of Guam establishments and rail transportation (NAICS 4821) establishments. Every quarter a national sampling frame list is created by combining all of the State lists into a single file called the Longitudinal Data Base (LDB). The following frame files were used to select the samples for the May 2004 survey.

- 2003 Second quarter LDB file (for May 2004 panel)
- 2002 Fourth quarter LDB file (for November 2003 panel)
- 2002 Second quarter LDB file (for May 2003 panel)
- 2001 Fourth quarter LDB file (for November 2002 panel)
- 2000 Fourth quarter LDB file (for 2001 sample)

In addition, the LDB files were supplemented with frame files covering Guam and rail transportation (NAICS 4821).

Stratification

Establishments on the frame are stratified by geographic area, industry group, and size class.

- Geographic stratification—686 MSA/BOS areas are specified within State. Each officially defined metropolitan area in a State is specified as a substate area. In addition, each State is allowed to specify up to four residual balance-of-State areas. (Note: Cross-State MSAs are split among the several States.)

- Industry stratification—343 NAICS4/5 industry groups are defined.

- Size class stratification—An establishment's size is defined as the maximum of its 12 monthly levels on the sampling frame. Establishments are classified into one of the following seven employment size class (SC) ranges: 1–4, 5–9, 10–19, 20–49, 50–99, 100–249, and 250 or more.

At any given time there are about 575,000 nonempty MSA/BOS-by-NAICS4/5-by-SC strata on the frame. When comparing nonempty strata between frames, there may be substantial frame-to-frame differences. The differences are due primarily to normal birth and death processes and to normal establishment growth or shrinkage. Other differences are due to NAICS reclassification and changes in geographic location.

Certainty and virtual certainty units

Federal Government and USPS units are are called "certainty" units because a census is taken of them every panel. A census is taken every other panel (covering November employment) of State government units. Technically, State units are not "certainty" units because data are not collected every panel. Consequently the term "virtual certainty" is applied to these units. The term "virtual certainty" also is applied to units employing 250 or more workers based on maximum employment (that is, units in size class 7). These units are sampled once in the six-panel survey cycle with certainty.

Allocation of the sample to strata

Each State is assigned a fixed overall sample size that is correlated to the size of its sampling frame. Nationally, a given frame has about 575,000 nonempty MSA/BOS-by-NAICS4/5-by-SC strata and most of the sample is allocated simply to get minimal coverage of all strata across a six-panel cycle. For each State, there also is a Neyman allocation procedure that is used to allocate the remaining sample to the strata. The objective of the Neyman allocation process is to even out the relative standard errors of estimates for "typical" occupations across all substate areas.

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 numbers allow us to minimize sample overlap between the OES survey and other large surveys conducted by BLS.

For each stratum, a specific PRN value is designated as the "starting" point to select a sample. From this "starting" point, we sequentially select the first "*n*" eligible establishments in the frame to include in the sample; "*n*" denotes the number of establishments to be sampled.

Panel weights (sampling weights)

Sampling weights are computed so that each panel roughly represents the entire universe of establishments.

Federal Government, USPS, and State government units are assigned a panel weight of 1. Other sampled establishments, including virtual certainty units in size class 7, are assigned a design-based panel weight. For a stratum with n establishments sampled from *N* frame establishments, the weight *N/n* is assigned to each of the n sampled establishments. *N/n* is the reciprocal of a frame establishment's chance of being selected into the panel.

National sample counts

The combined sample for the May 2004 survey is the equivalent of six panels. The sample allocations for the panels in this cycle are:

204,197 establishments for May 2004
204,881 establishments for November 2003
199,587 establishments for May 2003
201,016 establishments for November 2002
405,655 establishments for 2001 (two-panel equivalent)

In addition, some certainty units from 2000 were added to the sample to provide complete coverage of establishments with 250 or more employees. The combined sample size for the May 2004 estimates is approximately 1.2 million establishments, which includes only the most recent data for Federal and State Government. Federal and State Government units from older panels are deleted to avoid double-counting these industries.

Response and Nonresponse

Response

Of the approximately 1.2 million establishments in the combined initial sample, 1,103,140 were viable establishments (that is, establishments that are not out of scope or out of business). Of the viable establishments, 868,908 responded and 234,232 did not. The establishment response rate is 78.8 percent (868,908/1,103,140). The response rate in terms of weighted sample employment is 72.7 percent.

Nonresponse

Establishments that did not report occupational employment data are "unit" nonrespondents. Establishments that reported employment data but failed to break out employment by wage intervals are "partial" nonrespondents. Missing data for unit nonrespondents are imputed through a two-step imputation process. Missing data for partial nonrespondents are imputed through the second step of the process only.

Step 1, Impute an occupational employment staffing pattern.

For each unit nonrespondent, a staffing pattern is imputed using a nearest-neighbor "hot deck" imputation method. The procedure links a responding donor establishment to each nonrespondent. For the May 2004 survey, possible donors were respondents from the May 2004, November 2003, May 2003, and November 2002 panels. The nearest-neighbor hot deck procedure searches within defined cells for a donor that most closely resembles the nonrespondent by geographic area, industry, and employment size. The procedure initially searches for a donor whose reported employment is approximately the same as the nonrespondent's frame employment within the same MSA-BOS and five-digit NAICS. If the search is unsuccessful, the pool of donors is enlarged in incremental steps by expanding geographic area and industry until a suitable donor is found. Limits are placed on the number of times a donor can be used.

After a donor has been found, its occupational staffing pattern is used to prorate the nonrespondent's frame employment by occupation. The prorated employment is the nonrespondent's imputed occupational employment.

Note: At the end of step 1, total employment has been imputed by occupation for the nonrespondent. We do not, however, have an employment distribution across wage intervals for the occupations.

Step 2, Impute an employment distribution across wage intervals.

For each "unit" nonrespondent in step 1 or for each "partial" nonrespondent, impute an employment distribution across wage intervals for all occupations. This distribution, called the wage employment distribution, is imputed as follows:

- Identify the imputation cell for the nonrespondent. Imputation cells are initially defined by panel, MSA/BOS, NAICS4/5, and four size classes.
- Determine if the imputation cell has enough respondents to compute wage employment distributions. If not, incrementally enlarge the cell until there are enough respondents.
- For each occupation in the imputation cell, use the respondents to calculate an employment distribution as a percentage across wage intervals.
- Use the distributions above to prorate the nonrespondent's imputed occupational employment across wage intervals. (Or, for partial respondents, use the distributions above to prorate the reported occupational employment across wage intervals.)

Combining and benchmarking data for occupational employment estimates

Reweighting for the combined sample

Employment and wage rate estimates are computed using a rolling six-panel (3-year) sample. Estimates for the May 2004 survey were calculated using data from the May 2004, November 2003, May 2003, November 2002, and 2001 samples, plus a small number of virtual certainty units from the 2000 sample. Establishments from each sample were assigned weights to represent the universe at the time of sample selection. When the samples are combined, each sampled establishment is reweighted so that the aggregate sample represents the "recent" population. This revised weight is called the final weight.

In the reweighting process, only data from the most recent census of Federal Government, the USPS, and State government are kept. The final weight for establishments in these industries is set to 1. The final weight of virtual certainty units also is set to 1.

Noncertainty units were reweighted stratum by stratum. The original single-panel sampling weights were computed so that responses in a stratum could be weighted to represent the entire stratum population. In one common scenario, six panel samples are combined, and all six panels have sample units for a particular stratum. A summation of the single-panel weights would overrepresent the stratum population by a factor of six. Because we do not want to overrepresent the stratum population, the final weight of each establishment is set equal to one sixth of its single-panel sampling weight. In general, when six panel samples are combined, only *n* < 6 panels have sample units for a particular stratum. The final weight of each establishment in the stratum is computed by multiplying its single-panel sampling weight by 1/n.

Benchmarking to QCEW employment

A ratio estimator is used to calculate estimates of occupational employment. The auxiliary variable for the estimator is the average of the latest May and November employment totals from the BLS Quarterly Census of Employment and Wages (QCEW). For the May 2004 survey, the auxiliary variable is the average of May 2004 and November 2003 employment. To balance the State need for estimates at differing levels of geography and industry, the ratio estimation process is carried out through a series of four hierarchical employment ratio adjustments. The ratio adjustments also are known as benchmark factors (BMFs).

The first of the hierarchical benchmark factors is calculated in the States for cells defined by MSA-BOS, NAICS4/5, and employment size class (four size classes). If a first-level BMF is out of range, it is reset to a predetermined maximum ceiling or predetermined minimum floor value. First-level BMFs are calculated as follows:

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

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

Second-level BMFs are calculated for cells defined within States at the four-digit NAICS level by summing the product of final weight and first-level BMF for each establishment in the cell. Second-level BMFs account for the portion of universe employment that is not adequately covered by weighted employment in first-level benchmarking. Inadequate coverage occurs when "MSA-BOS | NAICS4/5 | size class" cells have no sample data or when a floor or ceiling is imposed on first-level BMFs. Second-stage benchmarks are calculated as follows:

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

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

Third-level BMFs ($\mathrm{BMF}_{3,\mathrm{H}}$) are calculated at the "State | 3-digit NAICS" cell level by summing the product of final weight, first-level BMF, and second-level BMF for each establishment in the cell. Fourth-level BMFs ($\mathrm{BMF}_{4,\mathrm{H}}$) are calculated at the "State | 2-digit NAICS" cell level by summing the product of final weight, first-level BMF, second-level BMF, and third-level BMF for each establishment in the cell. As with second-level BMFs, third- and fourth-level BMFs are computed to account for inadequate coverage of the universe employment.

A final benchmark factor, $\mathrm{BMF}_{\mathrm{i}}$, is calculated for each establishment as the product of its four hierarchical benchmark factors ($\mathrm{BMF}_{\mathrm{i}} = \mathrm{BMF}_1 * \mathrm{BMF}_2 * \mathrm{BMF}_3 * \mathrm{BMF}_4$). A benchmark weight value is then calculated as the product of the establishment's final weight and final benchmark factor.

Occupational employment estimates

Benchmark weights are used to compute estimates of occupational employment. Estimates are produced for cells defined by geographic area, industry group, and size of establishment (i.e., size class). To estimate the total employment for an occupation in a cell, take the product of reported occupational employment and benchmark weight for each establishment in the cell. Sum the product across all establishments in the cell. This sum is the estimate of total occupational employment in the cell.

The equation below is used to calculate occupational employment estimates for an estimation cell defined by geographic area, industry group, and size class.

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

o = occupation
h = estimation cell
w_i = benchmark 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 cell *h*

Wage rate estimation
Two externally derived parameters are used to calculate wage rate estimates. They are the following:

- Mean wage rates for each of the 12 wage intervals
- Wage updating factors (also known as aging factors)

Wage rates of workers are reported to the OES survey as grouped data across 12 consecutive, nonoverlapping wage bands. Individual wage rates are not collected.

An illustration. An establishment employs 10 secretaries at the following wage rates:
$ 7/hour - 1 secretary
$ 8/hour – 1 secretary
$11/hour – 2 secretaries
$12/hour – 2 secretaries
$13/hour – 2 secretaries
$15/hour – 1 secretary
$16/hour – 1 secretary

Wage rates for secretaries, however, are reported to the OES survey as follows:

Wage interval A (under $ 6.75/hour) – 0 secretaries
Wage interval B ($ 6.75-$8.49/hour) – 2 secretaries
Wage interval C ($ 8.50-$10.74/hour) – 0 secretaries
Wage interval D ($10.75-$13.49/hour) – 6 secretaries
Wage interval E ($13.50-$16.99/hour) – 2 secretaries
The remaining wage intervals have 0 secretaries.

Because wage rates are collected as grouped data, we must use grouped data formulae to calculate estimates of mean and percentile wage rates. Assumptions are made when using grouped data formulae. For the mean wage rate formula, we assume that we know the average wage rate for workers in each interval. For the percentile wage rate formula, we assume that workers are evenly distributed in each interval.

Wage data from panels in May 2004, November 2003, May 2003, November 2002, and 2001 were used to calculate May 2004 wage rate estimates. Wage data from different panels, however, are not equivalent in real-dollar terms. Consequently, wage data collected prior to the current survey reference period (May 2004) have to be updated or aged to approximate that period.

Determination of a mean wage rate for each interval
The mean hourly wage rate for all workers in any given wage interval cannot be computed using grouped data collected by the OES survey. This value is calculated externally using data from the Bureau's National Compensation Survey (NCS). Although smaller than the OES survey in terms of sample size, the NCS program, unlike OES, collects individual wage data. The mean hourly wage rate for interval L (the upper, open-ended wage interval) is calculated without wage data for pilots. This occupation is excluded because pilots work fewer hours than other occupations. Consequently their hourly wage rates are much higher.

Wage aging process
Aging factors are developed from the BLS Employment Cost Index (ECI) survey. The ECI survey measures the rate of change in compensation for nine major occupation groups on a quarterly basis. Aging factors are used to adjust OES wage data in past survey reference periods to the current survey reference period (May 2004).

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 was modified to use ECI aging factors.

$$\hat{R}_o = \frac{\sum_{z=t-5}^{t}\left(\sum_{i \in z} w_i \ BMF_i \ \hat{y}_{i\,o}\right)}{\hat{X}_o}$$

$$\hat{y}_{i\,o} = u_{z\,o}\sum_{r} x_{i\,o\,r}\,c_{z\,r} \qquad (i \in z)$$

o = occupation
$\hat{R}_o$ = mean hourly wage rate for occupation *o*
z = year (or panel)
t = current panel
w_i = final weight for establishment *i*
BMF_i = final benchmark factor applied to 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*
$x_{i\,o\,r}$ = 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)
$u_{z\,o}$ = ECI aging factor for panel (or year) *z* and occupation *o*
$c_{z\,r}$ = mean hourly wage for interval r in panel (or year) z

In this formula, $c_{z\,r}$ represents the mean hourly wage of interval *r* in panel (or year) *z*. The mean is determined externally using data from the NCS. Research is conducted periodically to verify the continued viability 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 mean or percentile hourly wage rate estimates by a "year-round, full time" figure of 2,080 hours (52 weeks x 40 hours) per year. These estimates, however, may not represent mean annual pay should 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 from annual wage rates with any reasonable degree of confidence.

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. *G* subsamples are created from the *G* random groups. Each subsample is reweighted to represent the universe.

G estimates of total occupational employment ($\hat{X}_{hjog}$), one estimate per subsample, are calculated. The variability among the *G* employment estimates is a good variance estimate for occupational employment. The two formulae below can be used to estimate the variance of occupational employment for an estimation cell defined by geographic area and industry group.

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

h = estimation cell defined by geographic area and industry group

j = employment size class (1-19, 20-49, 50-249, 250+)

o = occupation

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

G = number of random groups

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

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

$\bar{\hat{X}}_{hjo}$ = estimated mean employment for occupation *o* in cell *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 in cell *h* obtained by summing the variances $v(\hat{X}_{hjo})$ across all size classes *j* in cell *h*.

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

Occupational mean wage variance estimation

Because OES wage data are collected in intervals (grouped), they 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:

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

$\hat{R}_o$ = 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*

$\bar{q}_{ho}$ = 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*

$\left(\sigma_{cr}^2, \sigma_{er}^2, \textit{and } \sigma_{\omega r}^2\right)$ 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 will differ from the results of a census. An estimate based on a sample survey is subject to two types of error—sampling error and nonsampling error. An estimate based on a census is subject only to nonsampling error.

Nonsampling error

This type of error is attributable to 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 questions; 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.

Estimates of the sampling error for occupational employment and mean wage rate are provided in this publication to allow data users to determine if those statistics are reliable enough for their needs. Only a probability-based sample can be used to calculate estimates of sampling error. The formulae used to estimate OES variances are adaptations of formulae 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 yield, hypothetically, the true population value. In this case, the sample estimate and its standard error can be used to construct 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 were repeated many times, if each sample were surveyed under essentially the same unbiased conditions, and if an estimate and a suitable estimate of its standard error were made from each sample, the following would hold true:

- 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.

- 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.

- 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.

- 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 editing and quality-control procedures are used to reduce nonsampling error. For example, completed survey questionnaires are checked for data consistency. Followup 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.

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 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 as well as the time needed to review them.

Other quality control measures used in the OES survey include:

- Follow-up solicitations of nonrespondents, especially critical for 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

- Validation of the analytical tables of estimates at the NAICS4/5 level