B. Collection of Information Employing Statistical Methods

1. Description of Universe and Sample

a) Universe

Entity	Universe	Sample
Agricultural Region	12	12
Farm Labor Areas	498	90
Farms	2,000,000	1,008
Crop Workers (estimated)	1,800,000	3,000

The universe for the study is the population of field workers active in crop agriculture in the continental United States (U.S.). The National Agricultural Workers Survey (NAWS) will use multi-stage sampling relying on probabilities proportional to size to interview approximately 3,000 randomly selected crop workers in fiscal year (FY) 2009.

b) Response Rate

The sampling design (described below) involves obtaining a random selection of employers. In fiscal years 2002-2006, 68 percent of the randomly selected employers (or their surrogates) who employed workers the day they were contacted by interviewers agreed to cooperate in the survey and interviews were conducted on 58 percent of the eligible establishments. As there are no universe lists of workers, the sampling frame of workers is constructed after contact with the employer.

Once interviewers have a worker frame, a random sample of workers is chosen. The interviewers, who work in pairs, approach workers directly to set up interview appointments in their homes or other agreed-upon locations. Approximately ninety percent of the approached workers agree to be interviewed.

2. Statistical Methodology

Overview

The goal of the NAWS sampling methodology is to select a nationally representative, random sample of farm workers. The NAWS uses stratified multi-stage sampling to account for seasonal and regional fluctuations in the level of farm employment. The stratification includes three interviewing cycles per year and 12 geographic regions, resulting in 36 time-by-space strata. For each interviewing cycle, NAWS staff draws a random sample of locations within all 12 regions from a standing roster of multi-county sampling units. These county or multi-county units are the primary sampling units (PSUs). Growers within PSUs are the secondary level and workers within growers are the tertiary level of sampling units. The number of interviews allocated to each location is proportional to the crop activity at that time of the year. Interview allocation is thus proportional to stratum size.

In each location, a simple random sample of agricultural employers is drawn from a universe list compiled mainly from public agency records. NAWS interviewers then contact the sampled growers or farm labor contractors, arrange access to the work site, and draw a random sample of workers at the work site. Thus, the sample includes only farm workers actively employed in crop agriculture at the time of the interview.

Stratification

Interviewing cycles

To account for the seasonality of the industry, interviews are conducted three times a year in cycles lasting ten to twelve weeks. The cycles start in February, June and October. The number of interviews conducted in each cycle is proportional to the number of agricultural field workers hired at that time of the year. The U.S. Department of Agriculture's (USDA) National Agricultural Statistics Service (NASS) provides the Employment and Training Administration (ETA) the agricultural employment figures, which come from USDA's Farm Labor Survey. The NAWS visits a total of 120 interviewing locations each year. These locations are similarly apportioned among the cycles using NASS data.

Regions

Regional stratification entails defining 12 distinct agricultural regions that are based on USDA's 17 agricultural regions. At the start of the survey in 1988, the 17 regions were collapsed into 12 by combining those regions that were most similar e.g., Mountain I and Mountain II, based on statistical analysis of cropping patterns. In each cycle, all 12 agricultural regions are included in the sample. The number of interviews per region is proportional to the size of the seasonal farm labor force in that region, as determined by the NASS using information obtained from the Farm Labor Survey.

Sampling within Strata

Farm Labor Areas

Each region is composed of several multi-county sampling units called Farm Labor Areas (FLAs). Originally, the NAWS used USDA Crop Reporting Districts, but experience showed that these units were not homogeneous with respect to farm labor. As a result, using Census of Agriculture data and ETA mappings of seasonal farm labor concentrations, the NAWS staff identified aggregates of counties that had similar farm labor usage patterns and were roughly similar in size. The resulting FLAs also account for varying county size across the United States. For example, in the East, a Farm Labor Area may include several counties; in the West, a Farm Labor Area may be composed of a single agriculture-intensive county. FLA size is more homogeneous within region than it is across regions. There are 498 farm labor areas in the country and the NAWS has a standing roster of 90 FLAs which are chosen using probabilities proportional to size. These 90 FLAs contain 425 counties.

For each cycle, within each region, a sample of FLAs is drawn using probabilities proportional to size. The size measure used is an estimate of the amount of farm labor in the FLA during a particular cycle. In this case, the measure is based on the hired and contract labor expenses from the most recent Census of Agriculture (CoA), available at the time of drawing the sample. The CoA labor expenses are adjusted using seasonality estimates which identify the percentage of labor expenses that fall into each of the NAWS cycles: fall, spring and summer.

The seasonality estimates are constructed from Quarterly Census of Employment and Wages (QCEW) data. The estimates are made by aggregating the reported monthly employment for each month included in the corresponding NAWS cycle e.g., June, July, August, and September for the summer cycle. The percentage of employment corresponding to each cycle becomes that FLA's seasonality estimate.

An iterative sampling procedure is used to ensure an adequate number of FLAs is selected for each region. First, from a list of FLAs within region, a cumulative sum using the size of the seasonal hired and contract labor expenditures is constructed. The selection number is the product of a random number selected from the uniform distribution multiplied by the cumulative total of the seasonal hired and contract payroll. The FLA that includes that number in its selection of the cumulative sum is selected.

Example showing selection algorithm for FLAs within region

FLA	Seasonal labor expenditures	Cumulative Sum	Selected
A	100,000	100,000	
B	300,000	400,000	
C	800,000	1,200,000	
D	450,000	1,650,000	<=
E	600,000	2,250,000	

Selection

Random number	0.657
Random number * cumulative sum	1,477,205
Selected FLA	D

This single FLA selection is repeated (drawing one FLA at a time) until the sum of the seasonal labor expenditures for all the selected FLAs is equal to or exceeds a selected percentage. For example, FLAs are selected in all regions until the cumulative sum of their seasonal labor expenditures exceeds 40 percent of the region's total. The criteria number is a proportion sufficient to ensure that the number of FLAs selected meets the number of FLAs allocated for a cycle. The result is that the locations selected represent roughly the same proportion of farm labor expenditures in each cycle- region

combination. Interviews are allocated to each FLA proportional to the seasonal agricultural payroll.

Counties within FLAs are selected in a similar fashion. Counties are pulled one-at-a-time using a random point on an interval representing the cumulative sum of the seasonal labor expenditures for all counties within the FLA. This is done until counties representing 80 percent of the total labor in the FLA have been selected. Interviews begin in the first selected county and, as a county's work force is depleted, interviewing moves to the next randomly selected county on the list, until all the allocated interviews in that FLA have been completed. In FLAs where farm work is sparse, interviewers may need to travel to several counties to encounter sufficient workers to complete the FLA's allocation.

Employers

Within each selected county, employers are selected at random from a list of agricultural employers. The list is compiled from administrative lists of employers in crop agriculture. An important component of the list is employer names in selected North American Industrial Classification Codes that the Bureau of Labor Statistics (BLS) provides directly to the contractor per the terms of an interagency agreement between the ETA and the BLS. Because of uncertainty about the conditions of seasonal farm labor in each location, the number of employers to be interviewed is not known in advance. An algorithm ensures that a minimum of five employers are contacted in each FLA (See Appendix G, Part C, page 38). For example, if the county allocation is 25, up to five workers may be interviewed at each employer. If an employer has less than five workers, however, all of the workers may be interviewed. Interviewers follow the same procedure at the last employer selected, e.g. interviewing up to five or all of the workers, even if this causes the interviewer to exceed the county interview allocation.

Once the randomly selected employer is located, the interviewer determines if he/she is familiar with his/her work force. If not, the interviewer seeks the name of the packinghouse manager, personnel manager, farm labor contractor, or crew leader who can help construct a sampling frame of the workers in the operation. Interviewers document the number of workers employed on the day of worker selection in order to construct worker selection probabilities. The interviewers follow specific sampling instructions that were designed by a sampling statistician to ensure selection of a representative random sample of workers at each selected employer. These instructions are included in Appendix G.

Weighting

The NAWS uses a variety of weighting factors to construct weights for calculating unbiased population estimates:

- Sampling weights are used to calculate unbiased population estimates by assigning each sample member a weight corresponding to the inverse of its probability of selection.

- Non-response factors are used to correct sampling weights for deviations from the sampling plan, such as discrepancies in the number of interviews planned and collected in specific locations.
- Post-sampling adjustment factors are used to adjust the weights given to each interview in order to compute unbiased population estimates from the sample data.

As explained below, non-response weights are calculated simultaneously with regional post-sampling adjustment weights.

Sampling weights

Each worker in the sample has a known probability of selection. Information collected at each stage of sampling is used to construct the sampling weights.

Sampling weights are calculated as the inverse of the probability of being selected:

$Wt_i = 1/prob,$

where prob = workprob*growprob*counprob*flaprob,

with $workprob = \frac{\text{number of workers interviewed at the farm location}}{\text{total number of workers at that location}}$,

$$\text{growprob} = \frac{\text{number of growers interviewed in the county}}{\text{total number of qualified growers in that county}},$$

Calculating counprob, the county within FLA weight, and flaprob, the FLA within region weights, are more complicated. For example, if one of the sampled FLAs is larger than another, then its probability of selection should be higher than that of the other. If several FLAs are selected from a particular region, then the selection probability for a particular FLA is (1) its probability of selection on the first draw, plus (2) the probability of its selection on the second draw, plus (3) the probability of its selection on the third draw, etc.

For the standard method of sampling several items with probabilities proportional to size, without replacement, closed-form formulas for the exact inclusion probabilities do not exist. However, these probabilities can be calculated exactly using multiple summations. This procedure can be implemented in SAS within PROC IML.

Suppose that the population at a particular sampling stage consists of N objects with sizes $s_1, s_2, \cdots, s_N$, having total size $S = \sum_{j=1}^{N} s_j$. Let π_j^i be the probability that the j^{th} item is selected on the i^{th} draw. Then for $j = 1, 2, \cdots, N$,

$$\pi_j^1 = \frac{s_j}{S},$$

$$\pi_j^2 = \sum_{\substack{k=1 \\ k \neq j}}^{N} \frac{s_k}{S} \frac{s_j}{S - s_k},$$

$$\pi_j^3 = \sum_{\substack{k=1 \\ k\neq j}}^{N} \sum_{\substack{l=1 \\ l\neq j \\ l\neq k}}^{N} \frac{s_k}{S} \frac{s_l}{S - s_k} \frac{s_j}{S - s_k - s_l},$$

$$\pi_j^4 = \sum_{\substack{k=1 \\ k\neq j}}^{N} \sum_{\substack{l=1 \\ l\neq j \\ l\neq k}}^{N} \sum_{\substack{m=1 \\ m\neq j \\ m\neq k \\ m\neq l}}^{N} \frac{s_k}{S} \frac{s_l}{S - s_k} \frac{s_m}{S - s_k - s_l} \frac{s_j}{S - s_k - s_l - s_m}$$, and so forth.

These i^{th}-draw probabilities each have the property that $\sum_{j=1}^{N} \pi_j^i = 1$. Finally, the probability that the j^{th} item is included in a sample of size n is $\pi_j = \sum_{i=1}^{n} \pi_j^i$. These inclusion probabilities have the property that $\sum_{j=1}^{N} \pi_j = n$.

Both the FLA and county selection probabilities can be calculated exactly using these formulas.

Non-Response Weighting

Non-response corrections adjust for deviations from the sampling design. If, for example, ten interviews should have been collected at a farm but only two interviews were collected, those two interviews could be given five times the weight they would have received otherwise. Thus, each interview's weight needs to be adjusted to represent a certain value in terms of size.

Instead of making this adjustment at the farm level, it could be made at any higher level in the sampling plan. For the NAWS this means at the list-within-county, county, list-within-FLA, FLA, list-within-region, region, or national level.

By raising the level at which adjustments are made, overall size information is, generally, more reliable. This is due to the statistical effect of averaging, greater year-to-year stability over larger geographic areas, and the absence or suppression of data due to confidentiality considerations. On the other hand, lower-level adjustments are more sensitive, if the information used for making the adjustments is reasonably accurate.

For two reasons, the NAWS non-response adjustments are made at the region level. First, the region is the lowest level with enough interview coverage to calculate weights for the size adjustment. All of the 12 NAWS regions are visited in every cycle. If, for some reason, there are too few interviews in a region, the region can be combined with adjacent regions for weighting purposes.

Second, the NAWS uses measures of size provided by the USDA Farm Labor Survey, which are reported by quarter and region. The USDA is the only source of quarterly statistics on levels of farm worker employment. The Census of Agriculture, for instance,

reports data annually rather than quarterly and the statistics are published every five years. Thus, by using USDA Farm Labor Survey figures to make the size adjustment, the NAWS can adjust the weights by season and region and construct unbiased population estimates. Non-response adjustments for size, therefore, are made at the region-within-cycle level to create corrected region weights.

Post-sampling weights

Post-sampling weights are used in the NAWS to adjust the relative value of each interview in order for national estimates to be obtained from the sample. There are five post-sampling weights. Two of the weights adjust for unequal probabilities of selection that can only be determined after the interviews are conducted. These include the unequal probabilities of finding part-time versus full-time workers (day weight) and the unequal probabilities of finding seasonal versus year- round workers (seasonal weight). The next three weights (region, cycle, and year) adjust for the relative importance of a region's data, a sampling cycle, and a sampling year. The measures of size used are obtained from the USDA. The region weight, as discussed below, is calculated simultaneously with non-response weighting. The cycle weight and year weight serve slightly different roles in estimation. They allow different cycles and sampling years to be combined for statistical analysis. These weights are also based on USDA measures of size.

It should be noted that the NAWS sampling plan is based on USDA NASS data collected in the year before the interviews. For example, fiscal year 2008 data is used to plan the NAWS 2009 sample. The weights use NASS data from the year during which the interviews were conducted. This corrects for any discrepancies in allocations due to projecting farm worker distributions based on past year data.

Adjustment for days worked per week: the day weight

The day weight adjusts for the probability of finding part-time versus full-time farm workers. A part-time worker, who works only two or three days per week, has a lower likelihood of being encountered by the interviewing staff than a worker employed six days per week. Therefore, respondents are weighted inversely proportional to the length of their workweek.

A conservative adjustment for the number of days worked is appropriate to avoid excessively large sampling weights. Field reports indicate that relatively few workers are contacted on Sundays, and a review of the interviews indicated that virtually no workers reported Sunday hours without Saturday hours (only five of 1801 interviews in fiscal year 1989, or .3%). Accordingly, workers reporting at least six workdays per week nearly always have a full chance of selection. Thus, any workers reporting at least six days of work per week are treated as having a full chance of selection; adjustments are made only for those workers with less than six days of work per week. This choice of six (rather than seven) days affects the weights by less than 17 percent.

The day weight (DWTS) is computed as: DWTS = 6 / (length of the workweek) where "length of the workweek" is the number of days per week the respondent reports working at the time of the interview for the current farm task (if two tasks are reported, the one

with more days per week is used). Seven-day workweeks are truncated to six, as explained previously. For the few workers not reporting the number of days, DWTS is assigned a default value of 1.

Adjustment for seasons worked per year: the season weight

The calculation of worker-based weights is complicated by the fact that workers could, in general, be sampled several times a year. Furthermore, the USDA information does not provide a figure for the number of farm workers for the year. The USDA reports the number of farm workers working each season, so the same worker could be reported in multiple seasons. Because of this repetition of workers across seasons, it is impossible to derive the total number of persons working in agriculture during the year.

As employment information is not available for every worker for each quarter of the year, the only way to avoid double-counting of farm workers is to use the 12-month retrospective work history collected in the NAWS. Specifically, predicting future-period employment is achieved by imposing the assumption that workers who report having worked in a previous season would work in the next corresponding season. For example, a worker sampled in spring 2008 who reported working the previous summer (2007) is assumed to work in the following summer (2008). For some purposes, including the calculation of year-to-year work history changes, this assumption cannot not be used. For purposes such as obtaining demographic descriptions of the worker population, however, this assumption provides satisfactory estimates.

Further, it is assumed that a worker has an equal likelihood of being sampled in each season worked. This assumption is dependent on a balance between the amount of farm work done by the worker in each season and the number of interviews obtained in that region for the season. Recall that the NAWS interview allocation is proportional to county-level seasonal agricultural payroll. Thus, the probability of sampling is related to the amount of work performed by individual workers. With these simplifying assumptions, it is possible to calculate a seasonal weight that is simply the inverse of the number of seasons the interviewee did farm work during the previous year.

For the purposes of the NAWS, there are only three seasons per year. An interviewee always performed farm work during the trimester he\she was sampled. From the NAWS interview, it can be determined during which of the two previous trimesters the respondent also did farm work. If the interviewee only worked during the current trimester, the seasonal weight is 1/1 or 1.00. If the interviewee worked during the current trimester and only one of the two prior trimesters, the seasonal weight is 1/2 or 0.50. Finally, if the interviewee worked during the current and both of the prior trimesters, the seasonal weight is 1/3 or 0.33.

This season weight is similar to the day weight in the sense that respondents who spend more time (seasons) working in agriculture have a greater chance of being sampled. Therefore, the weighting has to be inversely proportional to the number of seasons worked in order to account for the unequal sampling probability.

The region weight
The region weight adjusts the relative weight of a region's data in relation to the number of interviews collected in that region. If the number of interviews collected was smaller than the regional allocation in the sampling plan, an adjustment weight greater than one is assigned to each interview in the region, and vice versa. These adjustments ensure that the population estimates are unbiased.

The region weight is based on USDA measures of regional farm employment activity. This is the best source of information available about farm workers. The USDA figures are reported by region and quarter, which allows the weight to be sensitive to seasonal fluctuations.

Correspondence between USDA data and the NAWS sampling cycles
The calculation of the region weight relies on two pieces of information: the USDA regional measures of size and the number of interviews completed in each region. The first step in the process of calculating the region weight is to apportion the USDA quarterly size figures among the three NAWS sampling cycles.

The USDA figures are reported quarterly. NAWS sampling years, however, cover non-overlapping 12-month periods (from September to August), which are divided into three cycles. Accordingly, it is necessary to adjust the USDA figures to fit the NAWS sampling frame by apportioning the four quarters into three cycles.

For example, the number of farm workers in the fall cycle for a region is assumed to be the total number of workers for that region in USDA quarter 1 of the current fiscal year (FY_c) plus one-third the number of workers for that region in USDA quarter 2 of the next fiscal year (FY_p). The formula for the winter, spring and summer cycles is constructed similarly.

Determining the NAWS region grouping according to interview coverage
The region weight is the size of a region divided by the total number of interviews in that region. Thus, if the number of interviews were to increase for the same size, the region weight would decrease. The region weight is attached to all interviews in the region.

The region weight (within cycle) is calculated as follow for each region j (1..nij) in cycle i:

$$PWTR_{ij} = \frac{USDA_{ij} / X_{ij}}{\sum_{j=1}^{nij} (USDA_{ij} / X_{ij})} * \frac{\sum_{j=1}^{nij} X_{ij}}{SDWTS_{ij}}$$

where USDAij is the USDA estimate for region j in cycle i,
Xij is the number of interviews for region j in cycle i,
DWTSij is the sum of farm workers day weights for region j in cycle i, 1<=DWTSk<=6 (k refers to a farm worker), so that SDWTSij=Xij if all farm workers in (ij) are working

full time and SDWTSij=6*Xij if all farm workers are working 1 day only a week in (ij).

Combining different sampling cycles: the cycle weight

The NAWS combines data from the different sampling cycles (seasons) within the same sampling year in order to generate more observations for statistical analysis. In order to combine cycles it is necessary to adjust for the number of farm work days represented in each cycle in relation to the number of interviews collected in the cycle. For instance, suppose the NAWS did not do proportional sampling as explained above but rather interviewed the same number of people in all three cycles in the 2007 fiscal year. If the USDA reported more workers for the fall and spring/summer cycles, as compared to the winter cycle, then the interviews in the fall and spring/summer would be worth relatively more in terms of size than the interviews conducted in the winter cycle. Accordingly, the interviews in the winter would have to be down-weighted in relation to the interviews in the other seasons (cycles) before the cycles could be combined.

The cycle weight is calculated similarly to the region weight, but at the cycle- rather than region- level. The sum of the USDA size for a cycle is divided by the number of interviews in that cycle. The cycle weight (or region weight within year) is calculated as follows for each region j (1..nij), cycle i in year Y:

$$PWTCR_{ij} = \frac{\frac{USDA_{ij}}{X_{ij}} * K_{ij}}{\sum_{i,j \in Y}\left(\frac{USDA_{ij}}{X_{ij}} * K_{ij}\right)} * \frac{\sum_{i,j \in Y} X_{ij}}{SSEADWTS_{ij}}$$

where $\mathrm{SSEADWTSij} = \sum_{k \in (i,j)} \mathrm{DWTS_k} * \mathrm{SEASWTS_k}$

$$K_{ij} = \frac{\sum_{k \in (i,j)} DWTS_k * SEASWTS_k}{\sum_{k \in (i,j} DWTS_k}$$

and

0.33<=SEASWTSk<=1 (k refers to a farm worker) and SEASWTSk=1 if the farm worker worked only one cycle during the year, so that if all farm workers for region j in cycle i worked full time and only one cycle in the corresponding year Kij=1 and SSEADWTSij=Xij

Combining different sampling years: the year weight

The year weight allows different sampling years to be combined for statistical analysis. It follows the same rationale as the cycle weight, but at the sampling-year level. If the same number of interviews are collected in each sampling year, those interviews taking place in years with more farm work activity are weighted more heavily in the combined sample.

Sampling years cannot be combined if the interviews are not comparable in terms of agricultural representation. In an extreme case, suppose that the NAWS budget tripled one of the sampling years, consequently tripling the number of interviews. If the two

sampling years were joined without adjustment, the larger sampling year would have an unduly large effect on the results.

To avoid this, the year weight is calculated as a ratio of the total number of farm workers in a sampling year to the number of interviews in that sampling year. The year weight (or region weight related to all years of interviews) is calculated as follow for each region j (1..nij), cycle i (the sum over i,j means all farm workers, all cycles all years):

$$PWTYCR_{ij} = \frac{\frac{USDA_{ij}}{X_{ij}} * K_{ij}}{\sum_{i,j}\left(\frac{USDA_{ij}}{X_{ij}} * K_{ij}\right)} * \frac{\sum_{i,j} X_{ij}}{SSEADWTS_{ij}}$$

with the same notations than for the preceding weights.

Obtaining the final weights

Once the individual weight components are calculated, final composite weights are calculated as the product of the day weight, the season weight, and region weight. The cycle and year are also factored into the composite weights when multiple cycles or sampling years are used. The composite weights are adjusted so the sum of the weights is equal to the total number of interviews at the next level of aggregation. These adjusted composite weights based on farm workers are then used for calculating the estimated proportion of workers with various attributes.

The individual observation weights are obtained at the farm worker level:

$$PWTRD_k = PWTR_{ij,k\in(ij)} * DWTS_k$$

This is the weight within cycle; it includes an adjustment for the length of the workweek but no seasonal adjustment.

$$PWTCRD_k = PWTCR_{ij,k\in(ij)} * DWTS_k * SEASWTS_k$$

This is the weight within a year; it includes both the length of the workweek and seasonal adjustment. This weight may be used for the analysis of one particular year of interview.

$$\mathrm{PWTYCRD_k} = \mathrm{PWTYCR_{ij,k\in(ij)}} * \mathrm{DWTS_k} * \mathrm{SEASWTS_k}$$

The composite weight (PWTYCRD) is used for almost all NAWS analysis. This weight allows merging several years of analysis together. It is included in the public access dataset.

3. Statistical Reliability

a) Response

Employer response
To maximize grower response, the contractor sends an advance letter to growers and provides them a brochure explaining the survey. The letter is signed by the survey director and includes the names of the interviewers and their contact information. For further information or questions, the letter and brochure direct growers to contact either the survey contractors (JBS International) at a toll free number or the Department of Labor's (DOL) Contracting Officer's Technical Representative (COTR). Grower calls are returned quickly. In addition, and before the start of every interview cycle, JBS provides the COTR a list of scheduled interview trips. The list includes the counties and states where interviews will be conducted, the names of the interviewers who will be visiting the selected counties, and the dates the interviewers will be in the selected counties. The COTR refers to the list whenever he receives a grower call to confirm the interviewers' association with the survey.

Both DOL and the contractor make presentations on the survey and provide survey information, e.g., questionnaires, to officials and organizations that work with agricultural employers. The NAWS has received the endorsement of several grower organizations. This improves the response rate since agricultural employers sometimes call their grower organization when considering survey participation.

Intensive and frequent interviewer training is also conducted as a means to increase employer response rates. Interviewers are trained in pitching the survey in various situations and, being well versed in the history, purposes, and use of the survey, are able to easily answer any questions or address any concerns an employer might have. In addition, when explaining the purpose of the survey to employers, interviewers clearly distinguish the survey from enforcement efforts by the Department of Homeland Security, DOL and other Federal agencies, and assure growers that their information is confidential.

Worker response
The survey's methodology has been adapted to maximize response from this hard-to-survey population. Interviewers pitch workers in English or Spanish, as necessary. All interviewers are bilingual and bicultural. In addition, interviewers make sure that potential respondents know that they are not associated with any enforcement agency, e.g., Immigration and Customs Enforcement. Interviewers explain the survey to workers and obtain their informed consent.

b) Non-response

The $20 honorarium to farm workers enables the survey to achieve an estimated worker response rate of 90 percent. This high level of response greatly aids in protecting the survey estimates from non-response bias. To reduce grower non-response, interviewers

are instructed to make several contact attempts at different times of the day and on different days of the week. Interviewer contact attempts are logged and the logs are monitored for compliance. When necessary, interviewers are instructed to accommodate a grower's preference for scheduling surveys and, if needed, the interviewer can request an extension of the field period.

To measure the effect of grower non-response on the survey's findings, the survey's statistician, project manger, and COTR are exploring the possibility of using the minimal information known about and/or collected from the non-cooperating growers, e.g., primary crop, county, number of workers employed, and quarterly hired farm payroll to generate proxy grower types. If it is possible to construct such proxies, then the demographic characteristics of workers employed on farms of cooperating growers of a particular type will be analyzed to determine if there are significant differences in the key demographic and employment characteristics of workers from participating vs. proxy non-participating growers.

c) Reliability

A probability sampling methodology will be used and estimates of the sampling errors will be calculated from the survey data.

Estimation procedure

1. At the highest level of the sampling design, the region/cycle level, stratified sampling was used. Sampling is then carried out at the lower levels, independently within each stratum.

The following description is excerpted from Obenauf[1]:

The stratified sampling technique divides the entire population into relatively homogenous groups that are mutually exclusive and exhaustive. Samples are then drawn from each of these groups (strata) by simple random sampling or an alternate method. The entire sample is a compilation of these independent samples from each of the strata. In stratified sampling, an estimate of the population mean can be made for each of the strata.
Estimate of population mean:

$$\bar{y}_{st} = \frac{\sum_{k=1}^{L} N_k \bar{y}_k}{N},$$

where N_k is the population size of stratum k and L is the number of strata into which the population is divided.
If a simple random sample is taken within each stratum (recall that other schemes can be used to draw a sample from each of the strata), the following represents an unbiased estimate of the variance of $\bar{y}_{st}$:

1 Obenauf, W. (2003), "An Application of Sampling Theory to a Large Federal Survey," Portland State University Department of Mathematics and Statistics.

$$Var(\bar{y}_{st}) \approx \sum_{k=1}^{L} \left(\frac{N_k}{N} \right)^2 \frac{s_k^2}{n_k} (1 - f_k).$$

The standard error of the estimator is the square root of this estimated variance, or

$$\text{S.E.}(\bar{y}_{st}) = \sqrt{\sum_{k=1}^{L} \left(\frac{N_k}{N} \right)^2 \frac{s_k^2}{n_k} (1 - f_k)}.$$

2. At the second stage of the sampling design, within each stratum, counties (or groups of counties) are treated as clusters.

The following description is another excerpt from ObenaufError: Reference source not found.

The population is again divided into exhaustive, mutually exclusive subgroups and samples are taken according to this grouping. Once the population has been appropriately divided into clusters, one or more clusters are selected … to comprise the sample. There are several methods of estimating the population mean for a cluster sample. The method most pertinent to this study is that involving cluster sampling proportional to size (PPS).

With PPS sampling, the probability (z_j) that a cluster *j* is chosen on a specific draw is given by $z_j = \frac{M_j}{M}$, where M_j is the size of the j^{th} cluster and *M* is the population size. An unbiased estimate of the population total is given by

$$\hat{y}_{pps} = \frac{1}{n} \sum_{j=1}^{n} \frac{y_j}{z_j} = \frac{M}{n} \sum_{j=1}^{n} \frac{y_j}{M_j} = M\bar{\bar{y}},$$

where y_j is the sample total for *y* in the j^{th} cluster, *n* is the number of clusters in the sample and $\bar{\bar{y}}$ represents the average of the cluster means.

To estimate the population mean, this estimate must be divided by *M*, the population size.

The variance of the estimator of the population total is given by

$$V(\hat{y}_{pps}) = \frac{M^2}{n} \left[\sum_{i=1}^{N} \frac{M_i}{M} \left(\frac{y_i}{M_i} - \bar{\bar{y}} \right)^2 \right],$$

This is estimated by $V(\hat{y}_{pps}) \approx \frac{M^2}{n} s_{mean}^2$, where s_{mean}^2 is the sample variance of the $\frac{y_j}{m_j}$ values.

For an estimate of the population mean,

$$\bar{y}_{pps} = \bar{\bar{y}} = \frac{1}{n} \sum_{j=1}^{n} \bar{y}_j = \frac{1}{n} \sum_{j=1}^{n} \frac{y_j}{m_j} \text{ and } V(\bar{y}_{pps}) \approx \frac{s_{mean}^2}{n}.$$

In two-stage cluster sampling, the estimated variance of the estimator is then given by an iterative formula:

$$Var(\bar{y}_{clus}) = E_1 \left[Var_2(\bar{y}_{clus}) \right] + Var_1 \left[E_2(\bar{y}_{clus}) \right].$$

This iterative formula is then generalized to compute the variance of the estimators in multi-stage sampling schemes with three or more levels. Exact formulas become intractable at this point, and the various statistical software packages rely upon either re-sampling methodology or linear approximations in order to estimate the variances and standard errors of the estimators.

The following is an excerpt from the SAS documentation for PROC SURVEYMEANS[2].

> The SURVEYMEANS procedure produces estimates of survey population means and totals from sample survey data. The procedure also produces variance estimates, confidence limits, and other descriptive statistics. When computing these estimates, the procedure takes into account the sample design used to select the survey sample. The sample design can be a complex survey sample design with stratification, clustering, and unequal weighting.
> PROC SURVEYMEANS uses the Taylor expansion method to estimate sampling errors of estimators based on complex sample designs. This method obtains a linear approximation for the estimator and then uses the variance estimate for this approximation to estimate the variance of the estimate itself (Woodruff 1971, Fuller 1975)[3,4].

SAS (e.g., Proc Surveymeans), allows the user to specify the details of the first two stages of a complex sampling plan. In the present case, the stratification and clustering at the first two levels are specified in Proc Surveymeans (strata region; cluster FLA). At the lower levels of the sampling scheme, the design attempts to mimic, as closely as is practical, simple random sampling. The software is not able to calculate exact standard errors, since it presumes true simple random sampling beyond the first two levels. The sampling weights will remedy any differences in selection probabilities, so that the estimators will be unbiased. The standard errors, however, are only approximate; the within-cluster variances at stages beyond the first two are assumed to be negligible.

In the "Surveymeans" procedure, the STRATA, CLUSTER, and WEIGHT statements are used to specify the variables containing the stratum identifiers, the cluster identifiers, and the variable containing the individual weights.

For the NAWS, the STRATA are defined as the cycle/region combinations used for the first level of sampling and coded in a variable called dmaregn. The CLUSTER statement contains the primary sampling unit, which is the FLA. The variable for FLA is county_cluster.

The WEIGHT statement references a variable that is for each observation i, the product of both the sampling weight Wt_i and the non-response weight $PWTYCRD_i$. This variable is called pwtycrd for historic reasons.

---

2 *SAS Institute Inc., SAS/STAT® User's Guide*, Version 8, Cary, NC: SAS Institute Inc., 1999, 61, 3.
3 Woodruff, R. S. (1971), "A Simple Method for Approximating the Variance of a Complicated Estimate," *Journal of the American Statistical Association*, 66, 411–414.
4 Fuller, W. A. (1975), "Regression Analysis for Sample Survey," *Sankhyā*, 37, Series C, Pt. 3, 117–132.

The Surveymeans procedure also allows for a finite population correction. This option is selected using the TOTAL option on the PROC statement. The total statement allows for the inclusion of the total number of PSUs in each strata. SAS then determines the number of PSUs selected per region from the data and then calculates the sampling rate. In cases such as the NAWS where the sampling rate is different for each strata, the TOTAL option includes a reference to a data set that contains information on all the strata and a variable _TOTAL_ that contains the total number of PSUs in that strata.

We include here sample code for Proc Surveymeans to calculated the standard errors for our key estimator WAGET1.

```
     .
proc surveymeans data=naws.crtdvars total=naws.regioninfo;
strata dmaregn;
cluster county_cluster;
var waget1;
weight pwtycrd;
```

Precision of key estimators

Two of the many variables of interest are FWRDAYS, which is the number of days worked per year by a respondent, and WAGET1, which is the average hourly wage of a respondent.

For the first of these variables, the number of days (FWRDAYS), the 2-standard-error confidence interval is 189 ± 15.94. That is, with approximately 95% confidence, the average number of days annually worked, per person, lies between 173.06 and 204.94. This constitutes a margin of error of $\pm 8.4\%$ of the estimated value.

For the second variable, the average wage (WAGET1), the interval is $\$7.99 \pm \$0.22$. With approximately 95% confidence, the average wage lies between $7.77 and $8.21. This yields a margin of error of ±2.75% of the estimated value.

There are numerous other variables of interest, whose standard errors vary greatly. These two are offered as examples that show some of the range of possible precisions obtained.

4. Tests

The questionnaire to be used in the survey was developed by the DOL with input from various Federal agencies. Except for the new questions on occupational mental health, place of birth of parents, child care arrangements, and alcohol consumption, the questionnaire will be unchanged from the version that OMB approved in the last submission. As such, the majority of the current questions have been used for twenty years, are well understood by the sampled respondents, and the data they provide are of high quality.

The new questions on occupational mental health and alcohol consumption have been used successfully in other surveys and have also been tested for understanding in the farm

worker population. The new place of birth questions are straight forward but were nonetheless tested on a small number of survey staff for flow and understandability. The new questions on child care arrangements were piloted and analysis of the data showed that the questions worked very well.

5. Statistical Consultation

The following individuals have been consulted on statistical aspects of the survey design: Stephen Reder and Robert Fountain, Professors, Portland State University, (503) 725-3999 and 503-725-5204; Phillip Martin, Professor, University of California at Davis (916) 752-1530; Jeff Perloff, Professor, University of California at Berkeley (510) 642-9574; and John Eltinge, the Bureau of Labor Statistics (BLS) (202) 691-7404.

The data will be collected under contract to the ETA by JBS International, Aguirre Division (650) 373-4900. Analysis of the data will be conducted by Daniel Carroll, ETA (202) 693-2795, and by JBS International, Aguirre Division.

Appendix A: Contacting and Selecting Farm Workers

A. A FARM WORKER QUALIFIES TO PARTICIPATE IN THE NAWS (ELIGIBLE), IF HE/SHE …

1. **WORKS IN** any type of crop agriculture in the United States. This includes "crops" produced in nurseries.
2. **WORKS IN** the production of plants or flowers (including work done in nurseries like planting, cultivating, fertilizing, grafting and seeding).
3. has worked in the last 15 days, at least 4 hours per day, for the contacted employer, and meets any of the criteria mentioned above.

B. A WORKER CANNOT PARTICIPATE IN THE NAWS (INELIGIBLE) IF HE or SHE:

1. Is related (husband, wife, or couple, son or daughter) to another person already interviewed in the same cycle for the same employer (member of the same family unit listed on the family grid in the NAWS questionnaire). Only one of them can be interviewed.
2. Was interviewed by NAWS within the last 12 months in the same location.
3. Is an "H-2A worker." H-2A is a program similar to the "braceros". An H-2A worker is a foreigner who is in the United States on a temporary work visa to work for a specific agricultural employer or association of agricultural employers for a specific period of time (less than a year). At the end of the period, the worker returns to his/her respective country.
4. Works exclusively with livestock (animals: such as bees, horses, fishes, pigs, cows, etc).
5. Hasn't worked for the contacted grower at least one day for 4 hours or more in the last 15 days.
6. Does "non-farm work" for the employer (mechanic, sales, office, etc).
7. Is a family member of the grower or employer and doesn't draw a salary like other farm workers.
8. Is the grower or employer or contractor.
9. Is a sharecropper that makes all operational decisions such as when, where and how to plant, harvest, etc.
10. Works for a packing house or cannery (packing or canning agricultural products) outside of the ranch. **Note:** Workers who are packers or caners can be eligible for the NAWS study if they satisfy the following two requisites:
    - **a)** the canning or packing plant is adjacent or located on the farm, **AND**
    - **b)** at least 50 percent of the produce being packed or canned originated from the ranch of the contacted grower.
11. Works for a landscaping company that just sells, installs, maintains or preserve trees or plants; this includes the planting of ornamental plants and placement of sod.

Whenever a worker doesn't qualify to participate, be gracious and thank him/her for their time and proceed to the next worker.

C. NUMBER OF INTERVIEWS PER GROWER

The Grower Lists indicates the total number of interviews allocated for your assigned county and the total number of interviews that must be completed for each employer. **NEVER** can the ***total county*** allocation be completed by interviewing workers from ***one single employer***.

The total number of interviews per employer is based on the following guidelines. If the county (or FLA) allocation is**…**

…fewer than 25 interviews, the maximum number of interviews allowed per grower is **5**
…from 26 to 40 interviews, the maximum number of interviews allowed per grower is **8**
…from 41 to 75 interviews, the maximum number of interviews allowed per grower is **10**
…more than 75 interviews, the maximum of interviews allowed per grower is **12.**

Note: If you are unable to complete the maximum number of interviews per grower, continue on to the next grower until you have reached your allocation, but never exceed the maximum allowed per grower. **For example**: if you have 24 interviews for your assigned county, and you only complete 3 interviews with one grower (maximum interview per grower is 5) because the grower has only 3 employees, continue with the next growers on the list until you complete the allocation, but never exceed more than 5 interviews per grower or employer. At the last employer complete the number of interviews allocated to growers in that county – EVEN IF YOU EXCEED THE COUNTY ALLOCATION.

D. LOCATING THE WORKERS

Once you get permission from the Grower/Employer (and you have documented the number of employed workers) ask the Grower/Employer where you can find the workers. If they are in different locations ask the Grower/Employer: "how many workers are in each location?" Also ask the Grower/Employer (or supervisor assigned by employer) for the best time and location to meet with them.

WORKERS' LOCATIONS

The best time to contact workers
Unless the Grower/Employer gives you permission to speak with his/her employees during working hours, do not make any contacts or appointments or try to interview the workers during their work hours.

Changing work locations
Once the Grower/Employer gives you permission to contact the workers, try to complete your contacts and interviews on the same day the grower gave you permission. You should be aware that from day to day it is common to find that workers in the field change location; and new workers can be in the same field on a different day.

The location of the field is not in the assigned county
If the location of the field or operation of the farm is located outside of the designated county, you **cannot interview** those workers. The farm workers must be physically working in the NAWS assigned county for the particular cycle. That is, it is not unusual that the same Grower/Employer may have farm land and workers in two different counties.

E. HOW TO CHOSE ELIGIBLE WORKERS FOR THE STUDY

Random Selection
As a sample of workers from a Grower/Employer is needed, the workers are to be chosen at random. All eligible workers of the Grower/Employer must have an equal chance of being chosen.

Workers in different areas (locations)
In the fields, it is common that people who have similar characteristics such as gender, age, birth place, type of work, ethnicity, and etc. tend to group together. If this is the case, you should randomly choose a proportional number of workers from each group or the sampling would not be a good representation. **For example:** for a certain Grower/Employer you have 2 crews of employees. One crew is comprised of single, males with an average age of 24 yrs old, and from Oaxaca with about 6 months of residency in the United States. In contrast, the second crew is comprised of single females. If you choose from only the first crew you will not have a good representation of that grower's employees.

Selecting workers located in different areas
If the Grower/Employer informs you that his employees are distributed over two fields (in the same county) use the proportional formula (below #4) to calculate how many from each field you need to interview. The same proportional formula should be used if you locate workers in different residencies. **For example**, if the workers live in two different labor camps or housing then find out how many live in each

dwelling and calculate proportionately how many you should interview from each dwelling.

Proportional selection of workers

When you find that workers are divided into different areas, randomly sampling from each group will be necessary to maintain equal likelihood of selection for everyone. The following formula serves as a guide to calculate the number of workers that should be selected when you find that workers are divided into different areas. In this example, there are 3 fields and you are allowed to conduct 12 interviews for this grower.

a	b	c
Number of workers per location	**Number of workers per location ÷ Total of workers**	**%X# total of interviews = 12**
Field A = 20	20 ÷ 30 = 66.6%	.666 **x** 12 = 08 interviews
Field B = 05	05 ÷ 30 = 16.6%	.166 **x** 12 = 02 interviews
Field C = 05	05 ÷ 30 = 16.6%	.166 **x** 12 = 02 interviews
Workers total = 30		**Total = 12 interviews**

Once you have determined the number of workers to be selected, identify the correct sampling interval. For example, If five workers are to be selected from a crew of 15, then select workers in intervals of three –every third worker. Count off the workers in order, e.g., from right to left or front to back, and select every third worker.

Appendix B: Response to Bureau of Labor Statistics Comments

Comment 1
Per Section II.4 of the attached checklist and explanatory materials, the OMB will expect to see justification of the estimates of burden-hours through, e.g., citation of technical reports or published studies that provide estimates of the number of minutes per questionnaire required in previous similar surveys.

Response to Comment 1
The following two paragraphs have been added to part 12 of the supporting statement:

The estimated average time of 57 minutes per questionnaire is based on twenty years of survey administration (the NAWS began in FY 1989) and is comparable to the average number of minutes per questionnaire required in previous similar surveys after accounting for differences in questionnaire content. In a 1997 survey of the demographic characteristics and occupational health of migrant Hispanic farm workers in six Northern California Migrant Family Housing Centers (McCurdy et al. 2003), in which 1,201 adult farm workers were interviewed in person several times over the harvest season, the University of California at Davis (UCD) authors reported that the initial questionnaire, available at http://mccurdy.ucdavis.edu/fwis/FW_ADULT_INIT.DOC, required approximately 30 to 40 minutes to complete.

The UCD questionnaire is similar to but shorter than the NAWS questionnaire. Like the NAWS questionnaire, it elicited demographic, employment, and health information. Unlike the NAWS, it did not include question domains on employment benefits, housing, asset ownership, participation in education and training programs, receipt of needs- and contribution-based social services such as welfare and unemployment insurance, occupational mental health, and child care services. In addition, the UCD questionnaire did not capture as much household demographic information as the NAWS.

Another survey similar to the NAWS was the California Agricultural Worker Health Survey (CAWS) http://www.cirsinc.org/SurveyInstruments.html . This survey was conducted in 1999 by the California Institute for Rural Studies, Inc., (Villarejo et al. 2000) http://www.calendow.org/uploadedFiles/suffering_in_silence.pdf . The main survey instrument, which borrowed generously from the NAWS questionnaire, and included a household grid and work grid that are essentially identical to those found in the NAWS, was administered in person to 971 California agricultural workers. The authors estimated that about 20 to 30 minutes were required to complete it. Unlike the NAWS, the CAWHS instrument included lengthy sections on access to health care services, self-reported health conditions and doctor-reported health conditions. Also unlike the NAWS, the CAWHS elicited health-related information about each member of the subject's household. These health sections comprised about 29 pages of the 70-page instrument. The CAWHS, however, did not include the occupational mental health and child care questions.

Comment 2

Per Section II.4 of the attached checklist and explanatory materials, the OMB will expect to see detailed justification of of the "total cost to the government" figure of $3 million. Also, per Section II.4,

> The OMB has stated that under the “total cost to the government” response to item A.14, the lead agency should include full cost of the entire study, including the cost of sample and questionnaire design; data collection; data management; editing, processing, analyzing, reviewing, and interpreting the information; preparation of reports and documentation; reviewing reports; and (if applicable) preparing public use files. If a contractor is doing some or all of the abovementioned tasks, then the cost of the contract should be included, but agency staff time devoted to managing or doing the project should be included as well. Some components, such as contractor costs, will be relatively easy to know precisely. Other components, such as the costs of government employee time, will need to be a good-faith estimate. The costs reported in A.14 should be broken down into major categories of expense.

Response to Comment 2
Part 14 of the supporting statement has been rewritten as follows:

The estimated total survey cost for FY 2009 is $3,108,375. This includes the cost of the contract ($2,981,054) and ETA employee time ($127,321). The contract costs include sampling ($170,861), questionnaire design and testing ($88,510), data collection ($2,566,513), and report and public data set preparation ($155,170).

Comment 3
I am not able to align the proposed weighting steps (pages 22-23) with the corresponding steps of the sample design presented on pages 20-22. Specifically:

Comment 3.a
Page 21 appears to indicate that within a given region, sampling of the FLAs will take place with probabilities proportional to size. Are you using pps systematic sampling, the Chao (1982, Biometrika) pps algorithm, or a different pps algorithm? (This has an impact on the second-order inclusion probabilities, and thus may have an impact on appropriate variance estimators.) In addition, under the proposed pps design, within a given region different sample FLAs would potentially have different selection probabilities, provided they have different sizes. However, page 23 appears to indicate that within a given region, all sample FLAs would have the same selection probability, which generally would not fit with a pps design.

Comment 3.b
For the "counprob" calculation, one would again here generally expect that within a given FLA, county-level selection probabilities generally would vary across counties, but this does not appear to match the current mathematical definition of "counprob" given in the proposal.

Response to Comments 3.a and 3.b

The identified section on FLA and county sampling has been rewritten as follows:

For each cycle, within each region, a sample of FLAs is drawn using probabilities proportional to size. The size measure used is an estimate of the amount of farm labor in the FLA during a particular cycle. In this case, the measure is based on the hired and contract labor expenses from the most recent Census of Agriculture (CoA), available at the time of drawing the sample. The CoA labor expenses are adjusted using seasonality estimates which identify the percentage of labor expenses that fall into each of the NAWS cycles: fall, spring and summer.

The seasonality estimates are constructed from Quarterly Census of Employment and Wages (QCEW) data. The estimates are made by aggregating the reported monthly employment for each month included in the corresponding NAWS cycle e.g., June, July, August, and September for the summer cycle. The percentage of employment corresponding to each cycle becomes that FLA's seasonality estimate.

An iterative sampling procedure is used to ensure an adequate number of FLAs is selected for each region. First, from a list of FLAs within region, a cumulative sum using the size of the seasonal hired and contract labor expenditures is constructed. The selection number is the product of a random number selected from the uniform distribution multiplied by the cumulative total of the seasonal hired and contract payroll. The FLA that includes that number in its selection of the cumulative sum is selected.

Example showing selection algorithm for FLAs within region

FLA	Seasonal labor expenditures	Cumulative Sum	Selected
A	100,000	100,000	
B	300,000	400,000	
C	800,000	1,200,000	
D	450,000	1,650,000	<=
E	600,000	2,250,000	

Selection

Random number	0.657
Random number * cumulative sum	1,477,205
Selected FLA	D

This single FLA selection is repeated (drawing one FLA at a time) until the sum of the seasonal labor expenditures for all the selected FLAs is equal to or exceeds a selected percentage. For example, FLAs are selected in all regions until the cumulative sum of their seasonal labor expenditures exceeds 40 percent of the region's total. The criteria number is a proportion sufficient to ensure that the number of FLAs selected meets the

number of FLAs allocated for a cycle. The result is that the locations selected represent roughly the same proportion of farm labor expenditures in each cycle- region combination. Interviews are allocated to each FLA proportional to the seasonal agricultural payroll.

Counties within FLAs are selected in a similar fashion. Counties are pulled one-at-a-time using a random point on an interval representing the cumulative sum of the seasonal labor expenditures for all counties within the FLA. This is done until counties representing 80 percent of the total labor in the FLA have been selected. Interviews begin in the first selected county and, as a county's work force is depleted, interviewing moves to the next randomly selected county on the list, until all the allocated interviews in that FLA have been completed. In FLAs where farm work is sparse, interviewers may need to travel to several counties to encounter sufficient workers to complete the FLA's allocation.

Sampling weights

Each worker in the sample has a known probability of selection. Information collected at each stage of sampling is used to construct the sampling weights.

Sampling weights are calculated as the inverse of the probability of being selected:

$Wt_i = 1/prob,$

where prob = workprob*growprob*counprob*flaprob,

with $workprob = \frac{\text{number of workers interviewed at the farm location}}{\text{total number of workers at that location}}$,

$\text{growprob} = \frac{\text{number of growers interviewed in the county}}{\text{total number of qualified growers in that county}}$,

Calculating counprob, the county within FLA weight, and flaprob, the FLA within region weights, are more complicated. For example, if one of the sampled FLAs is larger than another, then its probability of selection should be higher than that of the other. If several FLAs are selected from a particular region, then the selection probability for a particular FLA is (1) its probability of selection on the first draw, plus (2) the probability of its selection on the second draw, plus (3) the probability of its selection on the third draw, etc.

For the standard method of sampling several items with probabilities proportional to size, without replacement, closed-form formulas for the exact inclusion probabilities do not exist. However, these probabilities can be calculated exactly using multiple summations. This procedure can be implemented in SAS.

Suppose that the population at a particular sampling stage consists of *N* objects with sizes $s_1, s_2, \cdots, s_N$, having total size $S = \sum_{j=1}^{N} s_j$. Let π_j^i be the probability that the j^{th} item is selected on the i^{th} draw. Then for $j = 1, 2, \cdots, N$,

$$\pi_j^1 = \frac{s_j}{S},$$

$$\pi_j^2 = \sum_{\substack{k=1 \\ k \neq j}}^{N} \frac{s_k}{S} \frac{s_j}{S - s_k},$$

$$\pi_j^3 = \sum_{\substack{k=1 \\ k \neq j}}^{N} \sum_{\substack{l=1 \\ l \neq j \\ l \neq k}}^{N} \frac{s_k}{S} \frac{s_l}{S - s_k} \frac{s_j}{S - s_k - s_l},$$

$$\pi_j^4 = \sum_{\substack{k=1 \\ k \neq j}}^{N} \sum_{\substack{l=1 \\ l \neq j \\ l \neq k}}^{N} \sum_{\substack{m=1 \\ m \neq j \\ m \neq k \\ m \neq l}}^{N} \frac{s_k}{S} \frac{s_l}{S - s_k} \frac{s_m}{S - s_k - s_l} \frac{s_j}{S - s_k - s_l - s_m}$$, and so forth.

These i^{th}-draw probabilities each have the property that $\sum_{j=1}^{N} \pi_j^i = 1$. Finally, the probability that the j^{th} item is included in a sample of size n is $\pi_j = \sum_{i=1}^{n} \pi_j^i$. These inclusion probabilities have the property that $\sum_{j=1}^{N} \pi_j = n$.

Both the FLA and county selection probabilities can be calculated exactly using these formulas.

Comment 4

The point-estimation and variance-estimation notation and formulas given on pages 30-31 do not appear to match the weighting formulas given on preceding pages. In addition, some of the variance estimation formulas given on pages 30-31 appear to be based on single-stage cluster sampling, but the design presented on preceding pages appears to involve multistage sampling. The point-estimation and variance-estimation formulas on pages 30-31 would need to be aligned with the full design and weighting developed on preceding pages.

Response to Comment 4

The estimation procedure section now reads as follows:

Estimation procedure

1. At the highest level of the sampling design, the region/cycle level, stratified sampling was used. Sampling is then carried out at the lower levels, independently within each stratum.

The following description is excerpted from Obenauf[5]:

5 Obenauf, W. (2003), "An Application of Sampling Theory to a Large Federal Survey," Portland State University Department of Mathematics and Statistics.

The stratified sampling technique divides the entire population into relatively homogenous groups that are mutually exclusive and exhaustive. Samples are then drawn from each of these groups (strata) by simple random sampling or an alternate method. The entire sample is a compilation of these independent samples from each of the strata. In stratified sampling, an estimate of the population mean can be made for each of the strata.
Estimate of population mean:

$$\bar{y}_{st} = \frac{\sum_{k=1}^{L} N_k \bar{y}_k}{N},$$

where N_k is the population size of stratum k and L is the number of strata into which the population is divided.
If a simple random sample is taken within each stratum (recall that other schemes can be used to draw a sample from each of the strata), the following represents an unbiased estimate of the variance of $\bar{y}_{st}$:

$$Var(\bar{y}_{st}) \approx \sum_{k=1}^{L} \left(\frac{N_k}{N} \right)^2 \frac{s_k^2}{n_k} (1 - f_k).$$

The standard error of the estimator is the square root of this estimated variance, or

$$\text{S.E.}(\bar{y}_{st}) = \sqrt{\sum_{k=1}^{L} \left(\frac{N_k}{N} \right)^2 \frac{s_k^2}{n_k} (1 - f_k)}.$$

2. At the second stage of the sampling design, within each stratum, counties (or groups of counties) were treated as clusters.

The following description is another excerpt from ObenaufError: Reference source not found.

The population is again divided into exhaustive, mutually exclusive subgroups and samples are taken according to this grouping. Once the population has been appropriately divided into clusters, one or more clusters are selected … to comprise the sample. There are several methods of estimating the population mean for a cluster sample. The method most pertinent to this study is that involving cluster sampling proportional to size (PPS).
With PPS sampling, the probability (z_j) that a cluster j is chosen on a specific draw is given by $z_j = \frac{M_j}{M}$, where M_j is the size of the j^{th} cluster and M is the population size. An unbiased estimate of the population total is given by

$$\hat{y}_{pps} = \frac{1}{n} \sum_{j=1}^{n} \frac{y_j}{z_j} = \frac{M}{n} \sum_{j=1}^{n} \frac{y_j}{M_j} = M\bar{\bar{y}},$$

where y_j is the sample total for y in the j^{th} cluster, n is the number of clusters in the sample and $\bar{\bar{y}}$ represents the average of the cluster means.

To estimate the population mean, this estimate must be divided by *M*, the population size.

The variance of the estimator of the population total is given by

$$V(\hat{y}_{pps}) = \frac{M^2}{n}\left[\sum_{i=1}^{N}\frac{M_i}{M}\left(\frac{y_i}{M_i} - \bar{\bar{y}}\right)^2\right],$$

This is estimated by $V(\hat{y}_{pps}) \approx \frac{M^2}{n} s^2_{mean}$, where s^2_{mean} is the sample variance of the $\frac{y_j}{m_j}$ values.

For an estimate of the population mean,

$$\bar{y}_{pps} = \bar{\bar{y}} = \frac{1}{n}\sum_{j=1}^{n}\bar{y}_j = \frac{1}{n}\sum_{j=1}^{n}\frac{y_j}{m_j} \text{ and } V(\bar{y}_{pps}) \approx \frac{s^2_{mean}}{n}.$$

In two-stage cluster sampling, the estimated variance of the estimator is then given by an iterative formula:

$$Var(\bar{y}_{clus}) = E_1\left[Var_2(\bar{y}_{clus})\right] + Var_1\left[E_2(\bar{y}_{clus})\right].$$

This iterative formula is then generalized to compute the variance of the estimators in multi-stage sampling schemes with three or more levels. Exact formulas become intractable at this point, and the various statistical software packages rely upon either re-sampling methodology or linear approximations in order to estimate the variances and standard errors of the estimators.

The following is an excerpt from the SAS documentation for PROC SURVEYMEANS[6].

The SURVEYMEANS procedure produces estimates of survey population means and totals from sample survey data. The procedure also produces variance estimates, confidence limits, and other descriptive statistics. When computing these estimates, the procedure takes into account the sample design used to select the survey sample. The sample design can be a complex survey sample design with stratification, clustering, and unequal weighting. PROC SURVEYMEANS uses the Taylor expansion method to estimate sampling errors of estimators based on complex sample designs. This method obtains a linear approximation for the estimator and then uses the variance estimate for this approximation to estimate the variance of the estimate itself (Woodruff 1971, Fuller 1975)[7,8].

SAS (e.g., Proc Surveymeans), allows the user to specify the details of the first two stages of a complex sampling plan. In the present case, the stratification and clustering at the first two levels are specified in Proc Surveymeans (strata region; cluster FLA). At the lower levels of the sampling scheme, the design attempts to mimic, as closely as is practical, simple random sampling. The software is not able to calculate exact standard errors, since it presumes true simple random sampling beyond the first two levels. The

6 *SAS Institute Inc., SAS/STAT® User's Guide*, Version 8, Cary, NC: SAS Institute Inc., 1999, 61, 3.
7 Woodruff, R. S. (1971), "A Simple Method for Approximating the Variance of a Complicated Estimate," *Journal of the American Statistical Association*, 66, 411–414.
8 Fuller, W. A. (1975), "Regression Analysis for Sample Survey," *Sankhyā*, 37, Series C, Pt. 3, 117–132.

sampling weights will remedy any differences in selection probabilities, so that the estimators will be unbiased. The standard errors, however, are only approximate; the within-cluster variances at stages beyond the first two are assumed to be negligible. In the "Surveymeans" procedure, the STRATA, CLUSTER, and WEIGHT statements are used to specify the variables containing the stratum identifiers, the cluster identifiers, and the variable containing the individual weights. The PROC SURVEYMEANS statement has an option for a finite population correction that is selected as well.

For the NAWS, the STRATA are defined as the cycle/region combinations used for the first level of sampling and coded in a variable called dmaregn. The CLUSTER statement contains the primary sampling unit, which is the FLA. The variable for FLA is county_cluster.

The WEIGHT statement references a variable that is for each observation i, the product of both the sampling weight Wt_i and the non-response weight $PWTYCRD_i$. This variable is called pwtycrd for historic reasons.

The Surveymeans procedure also allows for a finite population correction. This option is selected using the TOTAL option on the PROC statement. The total statement allows for the inclusion of the total number of PSUs in each stratum. SAS then determines the number of PSUs selected per region from the data and then calculates the sampling rate. In cases such as the NAWS where the sampling rate is different for each stratum, the TOTAL option includes a reference to a data set that contains information on all the strata and a variable _TOTAL_ that contains the total number of PSUs in that stratum. In this case the variable total contains the number of total number FLAs (PSU) per region (stratum).

We include here sample code for Proc Surveymeans to calculate the standard errors for our key estimator WAGET1.

The relevant sections of the online SAS manual including the options available for Proc Surveymeans are included as Appendix C.

```
      .
proc surveymeans data=naws.crtdvars total=naws.regioninfo;
strata dmaregn;
cluster county_cluster;
var waget1;
weight pwtycrd;
```

Precision of key estimators

Two of the many variables of interest are FWRDAYS, which is the number of days worked per year by a respondent, and WAGET1, which is the average hourly wage of a respondent.

For the first of these variables, the number of days (FWRDAYS), the 2-standard-error confidence interval is 189 ± 15.94. That is, with approximately 95% confidence, the average number of days annually worked, per person, lies between 173.06 and 204.94.

This constitutes a margin of error of $\pm 8.4\%$ of the estimated value.

For the second variable, the average wage (WAGET1), the interval is $\$7.99 \pm \$0.22$. With approximately 95% confidence, the average wage lies between $7.77 and $8.21. This yields a margin of error of ±2.75% of the estimated value.

There are numerous other variables of interest, whose standard errors vary greatly. These two are offered as examples that show some of the range of possible precisions obtained.

Comment 5
Please provide the details of your calculation of the confidence bounds reported on page 32. Since the NAWS apparently will be moving from a quota-sampling design to a probability-sampling design, it would be especially important to provide mathematical details of the specific ways in which the confidence interval calculations have been adjusted to account for the potential change in design effects.

Response to Comment 5
In the following derivation, we will give an approximate comparison of the ratio of the standard error of estimators under the new strategy (which incorporates "probability sampling" at the final stage) to the standard error of estimators under the old strategy (which employed a type of quota sampling at the final stage). The design effect at all of the earlier stages of the sampling scheme would be unaffected by the effect at the final stage. The estimator of the population total and population mean are

$$\hat{t} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} \text{ and } \hat{\mu} = \frac{1}{N} \sum_{i=1}^{n} \frac{y_i}{\pi_i},$$

respectively, where π_i represents the selection probability of the i^{th} item in the sample (and $w_i = \frac{1}{\pi_i}$ is the weight of the i^{th} item). For the purpose of this derivation only, we will assume a constant variance throughout the sample, and negligible covariances within clusters, so that the variances and standard errors can be expressed in terms of $\sum_{i=1}^{n} \frac{1}{\pi_i^2}$.

We will consider and compare two situations. In the first situation, when the interviewer reaches the final grower, he will terminate his interviews upon reaching the preset total number of interviews for that county. In the second situation, the interviewer will continue interviewing the alloted number of workers from the final grower, even if this exceeds the preset county total. We will compute

$$\frac{\mathrm{SE}_2(\hat{t})}{\mathrm{SE}_1(\hat{t})} = \frac{\mathrm{SE}_2(\hat{\mu})}{\mathrm{SE}_1(\hat{\mu})} = \frac{\sqrt{\sum_{i=1}^{n_2} w_{2i}^2}}{\sqrt{\sum_{i=1}^{n_1} w_{1i}^2}} = \sqrt{\frac{S_2}{S_1}} = \sqrt{1 + \frac{\delta}{S_1}},$$

where $\delta = S_2 - S_1$ represents the change in the sum of the squared weights. The change in the squared weight of the i^{th} item can be written

$$w_{2i}^2 - w_{1i}^2 = w_{2i}^2\left(1 - \frac{w_{1i}^2}{w_{2i}^2}\right), \text{ so}$$

$$\frac{\delta}{S_1} = \frac{\sum_{i=1}^{n_2} w_{2i}^2\left(1 - \frac{w_{1i}^2}{w_{2i}^2}\right)}{\sum_{i=1}^{n_1} w_{1i}^2} = \frac{\sum_{i=1}^{n_2} w_{2i}^2\left(1 - \frac{\pi_{2i}^2}{\pi_{1i}^2}\right)}{\sum_{i=1}^{n_1} w_{1i}^2}.$$

Notice that this is, essentially, a weighted average of the terms of the form $1 - \frac{\pi_{2i}^2}{\pi_{1i}^2}$. For this derivation only, we will approximate this weighted average by the corresponding unweighted average. This is equivalent to presuming that the change in selection probabilities is independent of the weights at all of the prior sampling stages.

Using information from cycles 50 through 55, the sample size is $n_1 = 6907$. Under the revised sampling plan, an additional 217 workers would have been interviewed, so $n_2 = 7124$. There were 731 workers that were interviewed at the "last grower" in their county.

For the $6907 - 731 = 6176$ workers who were not interviewed at the "last grower", the only adjustment to their weights would be a scaling coefficient to account for the sample size change from 6907 to 7124. That is, for these 6176 workers,

$$1 - \frac{\pi_{2i}^2}{\pi_{1i}^2} = 1 - \left(\frac{6907}{7124}\right)^2 = 0.05999.$$

The selection probabilities for the remaining $731 + 217 = 948$ workers would increase by a factor of $\frac{948}{731}$ (and then the weights would be adjusted by the same scaling coefficient as above), so that, for these 948 workers,

$$1 - \frac{\pi_{2i}^2}{\pi_{1i}^2} = 1 - \left(\frac{948}{731}\right)^2\left(\frac{6907}{7124}\right)^2 = -0.58093.$$

Thus, the average for all 7124 workers is

$$\frac{6176(0.05999) + 948(-0.58093)}{7124} = -0.02530.$$

Finally, the ratio of standard errors of the estimators under the two sampling plans is approximately

$$\sqrt{1 + \frac{\delta}{S_1}} = \sqrt{1 - 0.02530} = 0.987.$$

That is, the standard errors might change by about 1.3%.

REFERENCES

Fuller, W.A. (1975), "Regression Analysis for Sample Survey," *Sankhyā*, 37, Series C, Pt. 3, 117-132.

Obenauf, W. An Application of Sampling Theory to a Large Federal Survey. Portland State University, Department of Mathematics and Statistics. 2003.

SAS Institute Inc., SAS/STAT® User's Guide, Version 8, Cary, NC: SAS Institute Inc., 1999, 61, 3.