STATISTICS IN MEDICINE
Statist. Med. 2000; **00**:1–6 *Prepared using* simauth.cls *[Version: 2002/09/18 v1.11]*

Prior enumeration: a proposal for enhanced sampling for farmworker surveys

Richard Mines[1], Coburn C Ward[2], Marc B. Schenker[3]

[1] *Private Consultant* [2] *San Joaquin General Hospital, Graduate Medical Education, Stockton, CA* [3] *University of California, Davis, Department of Public Health Sciences, Davis, CA*

SUMMARY

We discuss present approaches to sampling farmworkers and their tendencies to under-represent the poorest, least educated, and most socially displaced. We suggest a new approach, called Prior Enumeration which refers to the process of creating the sampling frame after clusters are sampled. After the sampling frame is complete, a second phase of sampling is carried out which may use strata information derived from the prior enumeration. Formulas for estimation are given. Copyright © 2000 John Wiley & Sons, Ltd.

1. Introduction

The objective of farmworker surveys is to produce information that allows service providers, policy and program designers and farmworker advocates to better fulfill their missions. The survey results should improve understanding of currently identified "problem areas" for farmworkers and identify new ones. Namely, individuals concerned with farmworkers need information not currently available that will lead to better designed approaches to solving the health and other social problems faced by farmworkers. This paper argues that a customized sampling approach designed for farmworkers is necessary to collect this information in an unbiased manner.

As with other hard-to-reach populations, it is difficult to define the population sampling frame and therefore obtain a complete and unbiased sample of the community. There is no pre-existing list of farmworkers from which to sample. This makes it difficult to achieve the goal of producing reliable information that could lead to improving the lives of all the subgroups of farmworkers. Much useful information can be collected in nonrandom work such as ethnographic research approaches based on network sampling techniques. However, probability-based sampling is crucial to many applications of program designers, policy makers

*Correspondence to: rkmines@volcano.net

Contract/grant sponsor: The National Institute for Occupational Safety and Health through The Western Center for Agricultural Health and Safety; contract/grant number: 2U50OH007-550

Contract/grant sponsor: The California Endowment; contract/grant number: 20043221

and advocates. This type of sampling has a built-in estimation procedure, forces better control of nonsampling problems, and allows for a random infrastructure for ancillary work done using nonrandom selection methods. The individuals concerned with farmworkers could benefit from random selection methods in order to carry out the following tasks:

1. Assess how to increase participation in U.S. institutions of an institutionally timid population.
2. Assess the needs of the subpopulations such as unaccompanied men.
3. Measure the impact of government programs like Medicaid, food stamps, and state programs targeting the poor on farmworkers.
4. Assess the makeup of households and the demographic composition at and associated with domiciles.
5. Assess health needs in order to design effective public health prevention programs.

Survey work on farmworkers has been carried out extensively in recent years. Progress has been accomplished in collecting probability samples among the population that can inform debates about policies and programs for farmworkers. However, an inherent bias in the survey work of under-representation of the most underprivileged among the farmworkers has not been removed entirely. These particularly disadvantaged individuals and households face uniquely prevalent problems, including tuberculosis, mental illness, child labor practices, minimum wage violations, illegally charging for rides and illegally charging for equipment. Such problems require distinct solutions, and survey techniques should be further honed to reduce bias to an absolute minimum. Therefore, a method that builds on past approaches is proposed to address the under-representation of important subgroups of farmworkers.

2. Methods

What is the bias? There is potential for sampling or selection bias in conducting research among farmworker populations, given the challenges of collecting accurate data among this group. Farmworkers live in a variety of housing situations. Most live in crowded housing, both detached houses and apartments. Since such a large proportion of those employed at farm work (over 60 percent) are solo or unaccompanied males (men here without their parents, wives or children), many farmworkers live as roomers with relatives in anchor families or crowded in "crash pads" of all men. Also, many live in garages, trailers and cars near where other farmworkers live. These conditions make it crucial that any survey of farmworkers based on a household sample is preceded by a thorough enumeration of the neighborhoods where farmworkers live. If the surveyor or census taker goes to the door or telephones the residence, the person who answers will most likely be a permanent resident living in the house or the most settled resident of the apartment. Unless approached properly, these people will not have any incentive to identify others living in or near their home, especially since the other occupants may be undocumented or living in violation of housing codes. The result of these sociological realities is that surveys not customized for farmworkers have an unavoidable bias against the poorest among them. These same excluded people may be those most neglected by social and medical services.

2.1. Current Approaches

There have been several attempts in recent years to correct the probable bias in official farmworker data. None of these methods include a way of checking the data generated against an extended universe as is contemplated in the approach outlined below. Still, there is a great deal of useful information gathered by these techniques, and they can be used to complement the approach we are recommending. First, it must be recognized that no approach can reasonably be expected to eliminate entirely the problem of under-representing the farmworker sub-populations that are the poorest, least settled and have the fewest connections to U.S. institutions. While efforts to minimize this bias have already been initiated by current approaches, this suggested approach of Prior Enumeration (PE) is a further step toward minimization. As will be explained below, PE includes a prior identification of all the dwellings in a geographic area and a prior enumeration of the people living in or associated with these dwellings

The three, already tested, sampling designs described below are the network approach, exemplified here by the Binational Farmworker Health Survey (BFHS), the employment-based method exemplified by the National Agricultural Workers Survey (NAWS), and the household sampling method exemplified by the California Agricultural Workers Health Survey (CAWHS). The PE approach described below is meant to be an improvement on the third, household method. All three surveys used in-person interviews with trained bilingual interviewers, familiar with the respondent population and with skill at obtaining the confidence of the respondents.

The BFHS was carried out in 1999-2000 in Mexico and the United States (Mines et al., 2001). The staff went to southern state of Zacatecas in Mexico and identified 10 villages that had heavy participation in U.S. farm work. Based on information collected from village elders, a universe list of every living person who had done U.S. farm work and was raised in the villages was created. The lists were cross-checked by other informants in the community. At the time of the survey, some of the individuals were in the home village while others could be found in settlement communities in the United States. A random selection was taken from the universe list for each village. The survey began in Mexico and after approximately 300 interviews were done in the village, the survey team moved to the United States where another 150 interviews were done. During the survey period in the village, the addresses and phone numbers of the randomly selected members of the community in the United States were collected from relatives and friends in Mexico. Since there was a high degree of concentration of emigres in just a few settlement areas from each village, it was possible to obtain a very high completion rate of the randomly selected individuals on both sides of the border at a reasonable cost.

There were several advantages to this selection approach. First, the interviewers became known in the networks of people among whom they were interviewing since they spent several months in one community network. The refusal rate was extremely low since most of the respondents had heard about the survey and its objectives before they actually were approached by the interviewer. Also, since the interviewers spent time in the home village, the rapport with the respondents was easy to establish rendering the collected information more accurate.

The BFHS was the most inclusive in terms of types of workers of any of the techniques since it included ex-farmworkers and people living on both sides of the border. It is a great advantage to collect information about ex-farmworkers since the reasons for leaving farm work and the long

term impacts of the work could be investigated. The healthy worker effect, a potential bias that can result when sick or injured farmworkers are selectively excluded from research, is minimized by obtaining data from former farmworkers. In addition, the transnational context for collecting the information made the collection more complete and more accurate in many ways than the competing approaches. First, the difficult-to-reach farmworkers give better information in the confines of their home village surrounded by their families and friends. Equally important, the sample is more inclusive if some of the interviews are done south of the border since individuals unlikely to cooperate north of the border are included. And, individuals that are out of the United States at the time of the survey that would ordinarily be excluded from the sample have an equal likelihood to be included as other eligible respondents.

The biggest disadvantage of the BFHS approach is the inability to generalize the results to a larger population. Although the selection within the villages is random, the choice of the villages was not representative of all parts of Mexico. Therefore, a large sample with many points of origin from many places in Mexico would be required to assure that the data can be generalized.

The NAWS is an on-going survey of farmworkers that has been conducted continuously three times a year by the U.S. Department of Labor since October 1988 (U.S. Department of Labor). Since it has conducted over 60,000 surveys since its inception, it is indisputably the best national sample of farmworkers ever collected. It has a multistage technique in which counties are chosen using a probability proportional to size approach based on payroll spent on farmworkers by county. Within the counties, grower lists are constructed from Bureau of Labor Statistics and Department of Agriculture employer inventories. The interviewers must follow a rigorous procedure in choosing the growers and then another procedure to select the workers employed by that grower in the chosen counties. The technique is similar to one used by Muhib and colleagues called venue-based sampling in which respondents are chosen at a common venue [8]. Theoretically, all employed farmworkers have a chance to be included in the sample.

The advantages of the NAWS are its huge size, the fact that it is a long-term time series data collection effort, and the potential completeness of the sample. The data from the survey show that the poor, the undocumented, and the solo males (unaccompanied by their nuclear family) are well represented in the survey. Another important advantage is the seasonality of the sampling. Since farm work is seasonal, workers engaged in only one season have an acceptable likelihood of inclusion.

The disadvantages derive from the unwillingness of some employers to collaborate. Since sampling is done after speaking to the employer who identifies where the workers can be located, uncooperative growers may bias the results. Another continuing challenge for the NAWS is to provide its interviewers with complete and accurate grower lists. The lists tend to be inaccurate and duplicative unless constantly updated and improved-an expensive process.

The third approach that has been piloted in recent years is the community-based household survey. The California Agricultural Worker's Health Survey (CAWHS) used a multistage sampling strategy focused on small farmworker towns [14]. One community was chosen randomly from each of the six major farm areas in California. An additional site was added in a purposive way to ensure that the San Joaquin Valley, where most California farmworkers live, had two sites. Teams from the California Institute for Rural Studies (CIRS) mapped all the dwelling units located within each community. This prior assessment of the sampled dwellings like the Prior Enumeration described below involved walking through the entire

geographic unit and visually locating every dwelling unit. Dwelling units were assigned unique identification numbers which were used to randomly select dwellings for enumeration. Each dwelling was visited by an interviewer, and if at least one eligible farmworker was present, then all the eligible farmworkers were enumerated at that time. One worker was chosen randomly from the dwelling for an interview. The CAWHS oversampled for women to make sure that enough female respondents were interviewed to allow adequate analysis.

The sampling advantages of the CAWHS approach result from its partial prior enumeration of the dwellings, though not of the dwellings' inhabitants. This approach allowed for fewer informal dwellings to be missed by the survey. As a result, the bias against the most disadvantaged population was greatly reduced in the CAWHS compared to more conventional surveys. However, as can be shown by comparing CAWHS results with data from the California NAWS survey, certain subgroups may have still been under-represented by the CAWHS survey. For example, the proportion of solo or unaccompanied males is much higher in the NAWS than in the CAWHS. Although the partial enumeration in the CAWHS was crucial for its improved performance, it did not enumerate the members of the households associated with the addresses with a separate visit prior to the interview phase of the survey. In the CAWHS, the enumeration occurred just prior to the survey during the same interview. This may have caused two problems. First, the sample had to be done in a hurry at the moment of the survey without an analysis of the population and a careful sampling from this universe provided by a full prior enumeration. Second, it is easier to obtain the full array of individuals associated with an address when that is the focus of the visit than when the interviewer is pressured to complete an interview during the same visit. It is likely that the persistent bias of under-representing the most disadvantaged (due to the timidity of this population) was present during the CAWHS implementation. Another problem with the CAWHS is the limitation to a few select small farmworker communities. Many, if not most, farmworkers live in farmworker neighborhoods of mid-sized towns and even cities in agricultural areas. By focusing on just a few towns, the efficiency for interviewers of finding farmworkers was increased, but many farmworkers living in larger towns were excluded.

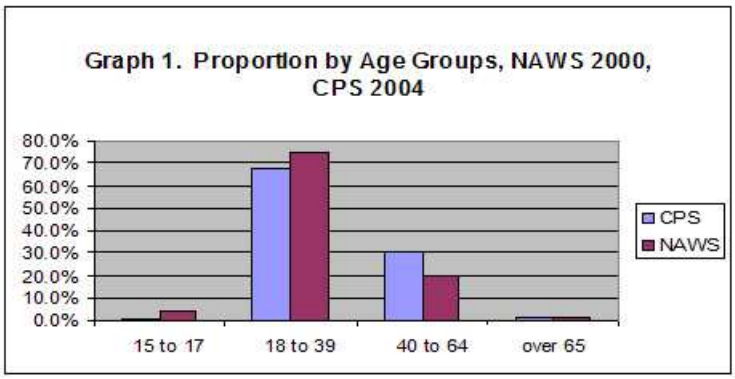
Though each of these approaches in its own way has contributed to the goal of gathering better farmworker data, none provides a method to check the data collected against a full or partial universe or sampling list. Even the NAWS with its huge sample size and continuous data collection can be checked only against employment estimates. The proposed PE approach suggests a method for checking demographic information collected in farmworker surveys against a larger universe.

3. Results

3.1. Demonstration of Bias

Although none of the existing farmworker-customized surveys have removed the bias completely, these surveys do minimize it. The reduction in bias can be demonstrated by comparing the respondents in these specially customized surveys to the farmworker respondents in mainstream surveys. These latter are often used as data sources to design farmworker programs or to discuss policies related to farmworkers. The bias manifests itself in an over-representation of the better-off, more established population farmworkers relative to the less

Figure 1.

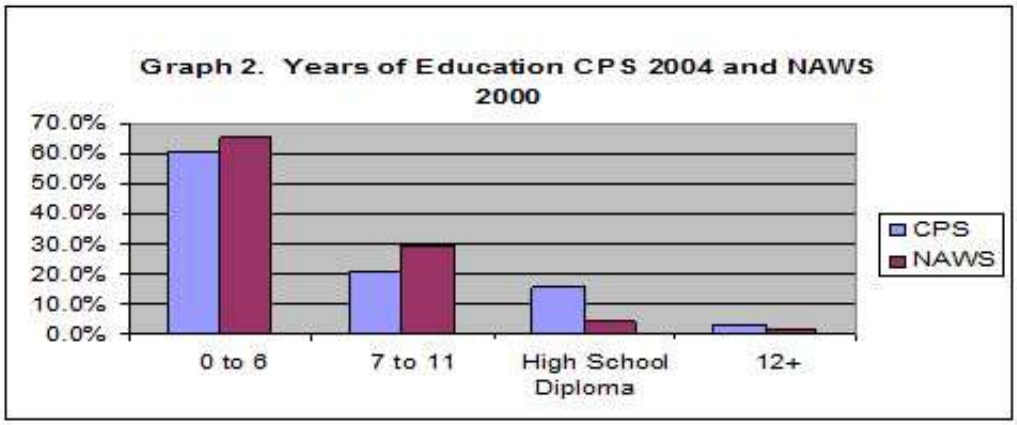


assimilated and more disenfranchised ones. It is precisely this bias that makes basing program planning on these non-specialized surveys so likely to lead to inappropriate program planning. It insures that the non-institutionalized remain marginalized.

First, we compare the Current Population Survey (CPS) annual summary of March, 2004 with the NAWS of 2000. Then California Health Interview Survey (CHIS) is compared to the CAWHS. Because of limitations on the availability of detailed occupational codes in the CPS, comparisons were made between two comparable foreign-born groups in the two surveys. All the foreign born in the CPS who worked in crop, livestock or in agricultural services were selected, with the assumption that almost all of the foreign born would be farmworkers. In the NAWS, only the foreign born were chosen since one must be a farmworker to be eligible for the NAWS.

The differences between the two groups of respondents are clear. NAWS respondents are younger, less educated and with lower incomes than CPS respondents (Mann-Whitney p ¡ 0.001 for each comparison). Proportionally more NAWS respondents are in the 18 to 39 range than CPS respondents (Figure 1). A randomly chosen person from the NAWS group would have a 56% chance of being younger than a randomly chosen person from the CPS group. There is a higher proportion of workers with 6 years or less education in the NAWS group and a smaller proportion with a high school diploma or more schooling. (Figure 2.) A randomly chosen person from the NAWS group has a 55% chance of having less education than a randomly chosen person from the CPS group. Finally, the majority of the CPS workers earn in the

Figure 2.



upper portions of the income ranges available to farmworkers while the NAWS workers mostly earn at the bottom of the scale (Figure 3). Fifty-seven percent of the NAWS respondents and 37% of the CPS respondents earn less than $12,500 per year while only nine percent of the NAWS respondents and 33% of the CPS respondents earn over $17,500. The estimated median (mean) yearly incomes are $15.6K ($17.9K) for CPS group and only $11.8K ($12.3K) for the NAWS. The CPS/NAWS comparison is limited by the aggregated nature of the CPS data available to the public. However, comparisons between the CAWHS and the CHIS more precisely identify the bias. The CAWHS sampling was done with a partial prior enumeration and in-person interviews at the home of the farmworker. The CHIS is a random selection of the entire California population and is done by telephone. Farmworkers in the CHIS were selected as a comparable group to the CAWHS. Both are household based health surveys that ask similar questions of the respondents.

The CHIS-sampled individuals appear to be more well connected and assimilated than the CAWHS respondents. In fact, the CHIS sample appears to greatly over-represent those of higher socioeconomic status (SES) while under-representing the less well off. The CAWHS by contrast, consistent with the NAWS, has a small proportion of higher SES workers. The CAWHS sample has 9.6% who finished high school, the CHIS sample 31.7%. The CAWHS sample has only five percent with annual income over $30,000 while the CHIS sample has 23%. In the CAWHS, six percent speak English well; in the CHIS, 16% speak it well. Six percent of CAWHS and 20% of CHIS participants were enrolled in MediCal. Likewise, 12% of CAWHS and 36% of CHIS sample were covered by employer-based insurance. Twenty-five percent of CAWHS and 61% of CHIS participants reported some type of health insurance. Among the CAWHS sample, only 17% of adults had dental visits in the last year, whereas in the CHIS sample, 48% reported a dental visit in the year prior to the survey. Among CAWHS participants, 47% are marginalized in the sense that, neither parents nor child have health insurance; however, in the CHIS in only 14% reported no health insurance for any member of the household.(All these differences are highly statistically significant: $p < 0.001$.) These

Figure 3.

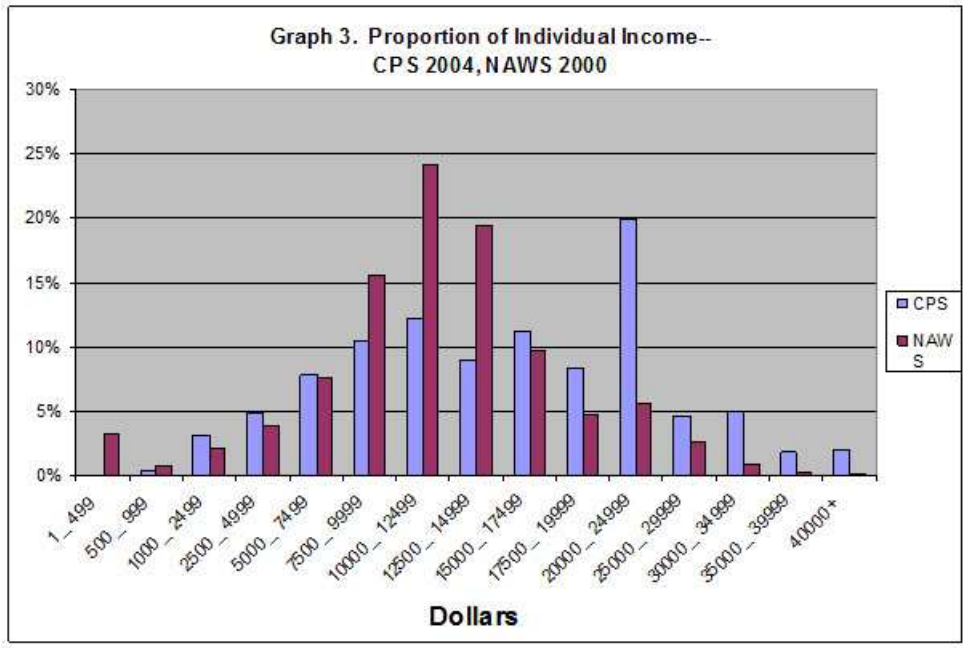


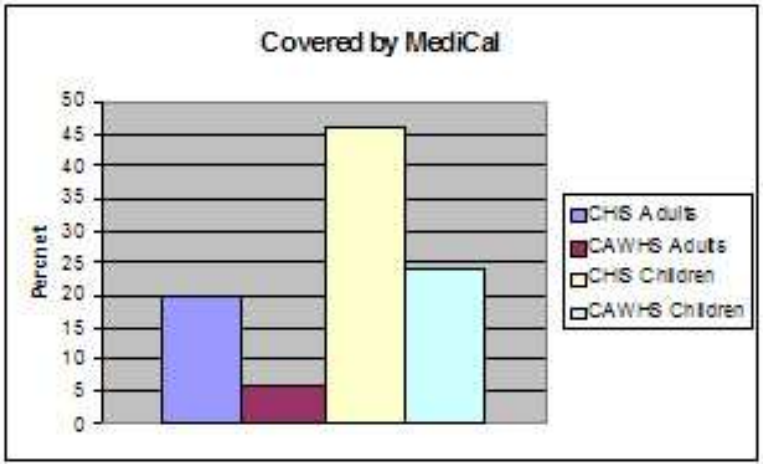
Table I. Children Covered by Insurance, by Birthplace

Survey	US Born	Mexico Born
CAWS	90%	47%
	n=20	n=303
CHIS	79%	84%
	n=14	n=181

comparisons suggest that the CHIS sample is a group much more able to access services and institutions.

Finally, despite the relative poverty of CAWHS respondents that would qualify them for MediCal use, more than twice as many of the CHIS adults and children are covered by the government insurance program. In fact, fully three quarters of the adults and half the children in the CAWHS are uninsured ($p < 0.001$). This latter finding is true even though the same percentage (82%) of the children in both surveys were born in the United States and are therefore eligible for government programs. U.S.-born children surveyed by CHIS are almost twice as likely (84% vs. 47%) to be covered by insurance than the similar group among the CAWHS respondents (Table I).

Figure 4.



4. Prior Enumeration (PE) of Domiciles in Farmworker Areas

4.1. A Description of the Approach

The first step in the process of PE is to choose an appropriate and feasible unit within which a database can be collected that fulfills the objective of the procedure. The objective is to create, at a reasonable cost, a database that will:

1. Identify and list members of all the subgroups of farmworkers. For example, subgroups would include both family-based units and solo male households. Farmworkers that live associated with addresses but who may not have a bedroom in which to sleep inside the house could be identified. This database will allow for the prior selection before the interview stage of individuals that represent in a statistically robust way all the subgroups in the population. For later sample selection at the time of survey implementation, the individuals must be identified in a recognizable fashion.
2. Permit the verification of the sample against a "representative" universe of farmworkers and the ex post weighting of the sampled population by reference to the universe.

Since the process described immediately below is labor intensive, the unit of PE data collection must be limited. Fortunately, most farmworkers live in a restricted number of U.S. Census tracts. The number of farmworkers identified in the 2000 Census in each Census Tract (about 1,000 people) can be obtained from the Census Bureau. In this way, a large geographic area such as a state or county could be surveyed by selecting a small group of eligible Tracts. Once the Tracts are chosen, the interviewers can begin the process of enumerating.

The first step of PE is to advertise in community fora that the survey is occurring so that people are not surprised by the appearance of interviewers. The next step is to review, by walking through the neighborhoods, every address in the tract, and identify all possible dwelling places and their associated addresses. In addition, areas without official addresses

Table II. Addresses by Married Couple or Solo Male: Mendota July, 2005

	Addresses (n)	Adults (n)	Children < 18(n)	Total Individuals
Solo Male	175	619	73	692
Married Couple	554	1821	994	2815
Total	729	2440	1067	3507

would be searched for identifiable dwelling locations. These dwelling locations might include garages, trailers, sheds and cars. Once all the potential dwelling places have been identified, in the second stage interviewers go to the addresses and determine if there are any farmworkers living at or associated with the address. Only addresses with at least one farmworker would be eligible for the survey. At the eligible addresses, the interviewers ask a series of simple questions about each person who is associated with one of the dwellings. The questions must be kept short and no names should be taken so as to maintain maximum trust at this point. The questions asked vary depending on the goals of the survey. Once the PE is complete, the survey staff can choose the sample including a stratified sample if deemed appropriate. Then, the implementation stage commences with pre-selected individuals (identified by their characteristics like age, gender, nationality) with predetermined backups for each address.

The procedure is best explained with an example. In the summer of 2005, a Prior Enumeration nearly identical to the recommended procedure was conducted in Mendota, California for a health survey of farmworkers. Census tracts and census blocks in Mendota were randomly selected. Then each address and some vacant lots with inhabitants were first reviewed for possible dwellings and then visited to collect the database. On a second stage of the PE, each address was visited and a series of questions were asked about each farmworker associated with the address. First, the relationship of the individual to the respondent (spouse, sibling, other relative, roomer) was determined; next the age, gender, place of birth, years of presence in Mendota, and whether the person had done two or more weeks of U.S. farm work in the prior year was asked for all those over 17 associated with the address. Finally, the number of minors under 18 at the address was asked. These data, though very limited in scope and without names, allowed for a careful review of the population for sampling purposes. Also, this provided sufficient information for verifying the representativeness of the sample against the total enumerated population.

The results allowed for a preliminary analysis of farmworker community traits that are very useful for selecting strata to sample. For example, by manipulating the few data points that were acquired, we were able to get a clear picture of the distribution of solo male and married couple-based households (Table II). In addition, we could identify the relationship of solo males living in married couple households to the household head—often they were siblings, cousins or other relatives.[†]

[†]The Census has a question about the occupation of the respondents. Those who are in NAICS codes 111 (crop), 112 (livestock) and 115 (agricultural services) and have a SOC code of farmworker 45-20xx could be used as a proxy for the farmworker population. A probability proportional size method is described below that will eliminate most census tracts from the PE. To order a special run look at http://www.census.gov/population/www/cen2000/sptabs/main.html.

Table III. Traits by Married Couple and Solo Male Addresses-Mendota July, 2005

Characteristic		
Married	Couple Address	Solo Address
Male > 18 years	61%	80%
Central American	29%	50%
Mean Years Mendota Resident	10 yrs	5 yrs
Mean age	33 yrs	30 yrs
Children < 18 years	35.30%	10.5%

Also, addresses could be analyzed in many other ways that help in stratifying the sample and in weighting the sample after the data is collected. Table III presents comparisons of the population in terms of percent male, percent Central American, mean time spent in Mendota, the mean age, and the number of minor children by the type of address (solo male or married couple). This level of understanding of the population prior to sampling is crucial for obtaining and checking to see if one has a representative sample.

In sum, the PE allows for the careful choice of whom to sample including the use of stratified sampling. It also allows for ex-post weighting of the sample using the total enumerated population as a universe of all those farmworkers in the Census Tracts chosen for study.

5. Sampling Scheme

The proposed sampling scheme consists of two stages with the prior enumeration taken between the two stages. Stage one is a cluster sampling with clusters being the census blocks in the area under study. The census block is a convenient unit for which the Census Bureau can estimate the number of farmworkers. The second stage is accomplished with a stratified sample drawn within each of the chosen clusters.

In the first stage, for purposes of cost efficiency, it is suggested to limit the possible selections to only those blocks with a minimum number of farmworkers. This estimated number can be acquired by a special run done by the Census Bureau of the most recent decennial census. Due to the concentrated nature of farmworker populations, this process will eliminate the vast majority of census blocks except in heavily farmworker neighborhoods. This resource-efficient limitation will introduce a degree of invalidity whenever conclusions are drawn about a whole group of farmworkers which covers those in the census block that were excluded a priori. However, the error is small if the procedure excludes only a very small proportion of farmworkers. Alternatively, the error will be small if there is no strong statistical association between variables of the study and whether there are few or many farmworkers in a given census block. In the subsequent discussion "population" will refer to all farmworkers in all the census blocks that were deemed eligible for selection at stage 1.

Among the N eligible clusters, n of them are chosen by a probability proportional to size (with replacement) sampling scheme. If M_i is the size of the farmworker population in cluster i and M is the total number of farmworkers in the "population", then clusters are chosen in sequence so that on each draw cluster i has a probability of being chosen with probability $p_i = M_i/M$, independent of what has been chosen on previous draws. This scheme allows

Table IV. Notation

Notation	Definition	How (when) Evaluated
N	Number of eligible clusters	Census Bureau, Design Decision
n	Number of clusters sampled	Design Decision
M	Total Farmworkers in Eligible clusters	Census Bureau
M_i	Farmworkers in i-th chosen cluster	Census Bureau
M_{ij}	Farmworkers in stratum j of i-th chosen cluster	Result of PE
m_{ij}	Sample size in stratum j of i-th chosen cluster	Design Decision
η_i	Number of strata in i-th chosen cluster	Design Decision
$\overline{y}_{ij}$	Sample average in stratum j of i-th chosen cluster	Stage 2 Result
N_j	Farmworkers in stratum j, totaled over all chosen clusters	Result of PE
$\overline{y}_i$	Estimate of mean of y within the i-th chosen cluster	Stage 2 Result

that an individual cluster might be chosen more than once; in such a case, say if one has been chosen $t > 1$ times, the second stage sampling is done t independent times within that cluster.

The second stage scheme involves stratified sampling within each chosen census block. Before the second stage of sampling occurs, the PE is completed (only once) in each cluster that is chosen. This gives a good sampling frame for the second stage of sampling. The definitions of the strata will be partially determined by previous knowledge of farmworker traits but will be enriched by the information collected in the PE database. Previous knowledge of farmworker survey results may be adequate to define strata and sample sizes that fit with the purpose of each survey. However, the PE will give precise information about the possible strata and their relative sizes for the specific universe to be sampled. Researchers would be free to choose sample size of strata in the second stage. In the case of a stratum composed of hard-to-find interviewees a large sample size could be assigned.

In order to clarify the notation to follow, Table IV gives the definitions and stages of instantiation.

Suppose one wants to estimate μ , the mean of a variable y over the whole "population". An example of this would be if y is age and μ is the mean age in the population. Let $\overline{y}_i$ denote the (unbiased) stratified sampling estimate of the mean of variable y within the i-th chosen cluster. The estimator of μ , denoted $\widehat{\mu}$, is given by

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \overline{y}_i \quad .$$

In the example, the estimate of μ is the simple average of the within cluster estimated means - the unexpected simplicity of the formula is due to sampling with probability proportional to size. The variance of $\widehat{\mu}$ is estimated by

$$\widehat{var}(\widehat{\mu}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\overline{y}_i - \widehat{\mu})^2.$$

The standard formula for calculating $\overline{y}_i$, the estimate of the mean of variable y within the

i-th chosen cluster, is

$$\overline{y}_i = \frac{1}{M_i} \sum_{j=1}^{\eta_i} M_{ij} \overline{y}_{ij}.$$

In the above formula M_{ij} represents the size of stratum j in the i-th chosen cluster, η_i the number of strata in the i-th chosen cluster (usually taken to be constant over i) and $\overline{y}_{ij}$ represents the sample average of sampled values in stratum j of the i-th chosen cluster. The estimator $\overline{y}_i$ has a variance estimated by

$$\widehat{var}(\overline{y}_i) = \sum_{j=1}^{\eta_i} \left(\frac{M_{ij}}{M_i} \right)^2 \left(\frac{M_{ij} - m_{ij}}{M_{ij}} \right) \frac{s_{ij}^2}{m_{ij}}$$

where M_{ij}, m_{ij}, and s_{ij}^2 respectively represent the number of farmworkers, the sampled number of them, and the sample variance of the values of y (all in stratum j of cluster i.)

5.1. Stratum Estimates

Often the researcher is interested in the mean value of y within a single stratum. For example, the mean income of the stratum of unmarried males can be estimated.

To estimate the mean of variable y in stratum j of the "population", we use the estimator $\widehat{\mu}_{\bullet j}$ given by

$$\widehat{\mu}_{\bullet j} = \sum_{i=1}^{n} \overline{y}_{ij} \frac{M_{ij}}{M_i}$$

where $N_j = \sum_{i=1}^{N} M_{ij}$ is the number of farmworkers in stratum j of the "population". The estimated variance of $\widehat{\mu}_{\bullet j}$ is

$$\widehat{var}(\widehat{\mu}_{\bullet j}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{\overline{y}_{ij} M_{ij} M}{M_i N_j} - \widehat{\mu}_{\bullet j} \right)^2,$$

the notation being described earlier.

Both the overall and the within stratum mean estimators are unbiased. Confidence intervals are calculated by using t multipliers with $n-1$ degrees of freedom.

If one wants to estimate percentages, then these formulas apply with one change: replace s_{ij}^2 by $\frac{p_{ij}(1-p_{ij})}{m_{ij}-1}$ where p_{ij} represents the sampled proportion of individuals in stratum j in the i-th chosen cluster with the characteristic to be measured. Also adjustments can be made so that whenever a cluster has been selected $t > 1$ times before prior enumeration, we can take a single sample of tm_i farmworkers without replacement rather than t independent samples of size m_i . This reduces the standard error of the estimator.

6. Discussion

6.1. Evaluating Prior Enumeration

Area sampling is a long-used procedure, and methods of choosing to sample the appropriate areas among many geographic areas in the universe similar to the PE approach have been

successfully implemented by many survey organizations. It has been used by the National Opinion Research Center and the Research Triangle Institute to perform national surveys for health expenditures for the Federal Government [4]. Also, the U.S. Department of Agriculture's National Agricultural Statistical Service uses an area sampling design (i.e. choosing a random sample of areas) to pick respondents in its Quarterly Agricultural Labor Survey. The NAWS also uses area sampling; it picks counties with a probability proportional to size technique that chooses by farm payroll reported in the Agricultural Census.

One crucial advantage of the PE is that it allows for the reduction of nonsampling errors. The interviewers' prior visit to the addresses that establishes the composition of the members of the households living at the address allows for sampling prior to the survey implementation stage. The interviewer can enter the premises and request to speak to the proper respondent without a cumbersome and distracting choice of respondent after arriving. Also, familiarity with the nationality, age, and relationships of residents at the address can allow for better rapport between the interviewer and the members of the household (see F. Floyd in [1] p. 259). Another nonsampling advantage of PE is that the questions included in the questionnaire can be pre-designed for the demographic traits of the person to be interviewed. This allows for the avoidance of skip patterns and the expansion of questions customized for one stratum of workers. Also, prior knowledge of the characteristics of the interviewees can be utilized during the training of interviewers. Interviewers can be trained to specialize in certain types of respondents and given special lessons on how to gather information from these individuals. For example, backup prompts that are designed to elicit responses for questions that interviewees may not comprehend can be tailored for the different demographic groups without jeopardizing the standardization of possible responses. Since the type of respondent at each address is known prior to the survey, the interviewers can be assigned to interview predominantly respondents for whom they have received customized training [1](see Fowler, Chapter 14, Biemer, et al).

There are advantages also in reducing sampling errors by using the PE approach with farmworkers. Underrepresentation is typically associated with homelessness and street people [4]. But, among farmworkers, overcrowded and ancillary dwelling units can cause the same problems. The PE approach allows for the sampling among the most disadvantaged groups that usually are greatly undercovered by traditional population surveys. The PE approach allows for the identification and selection of difficult-to-reach individuals without eliciting names allowing them to be sampled and found. Further, prior knowledge of the probability of selection of (almost all) the universe elements in the chosen census blocks can lead to accurate measurement of undercoverage of subgroups. The nonresponse to certain questions can also be monitored by subgroup more carefully. For example, Hispanics are known to avoid answering questions about social relations. [9]

Also, the PE improves the ability to correctly stratify the sample. The near complete coverage of certain limited demographic traits of the universe reduces any misclassification and improper assignment of individuals to strata. The PE can also integrate "insider information" gathered informally in the community to probe for missing individuals to achieve as complete an enumeration as possible. The PE allows for the oversampling of certain small or hard to reach strata and the use of post survey weights to adjust the results. It also allows for weighting of undercovered strata or subgroups. Survey results can be best analyzed if they can be checked "so that estimates agree with existing parameters" ([4] Chapter 2). However, without a PE database, there are no reliable benchmarks for farmworkers.

Another strength of the PE approach is the establishment of a random sample infrastructure

that can be utilized in interpreting and situating the work of complementary nonrandom network- based interviewee selection projects. Further ethnographic work can be done pursuing questions unresolvable by quantitative surveys among certain subgroups of the population. The significance of the findings of this nonrandom work can then be put into perspective by situating the group analyzed in the total universe of farmworkers. For example, problem networks like affinity groups with high rates of diabetes can be studied and then the importance of the findings can be quantified by reference to the universe of farmworkers gathered by the PE database. Finally, the PE allows for the analysis of the data by various units of inquiry depending on the goals of the research. The individual, nuclear family, household and residents of a given address can all be chosen as units of analysis with access to the PE database.

6.2. Conclusions

The improvement of farmworker survey data is possible that could include various methodological approaches. The need to maintain a probability based structure under this improvement is crucial to keep the data as useful to service deliverers, advocates and policy makers. The PE approach allows for a framework to reduce the bias against the most disadvantaged, to allow for systematic complementary research using nonrandom methods, and for the maximum use of post-survey weighting to obtain statistically reliable information.

REFERENCES

1. Biemer P, Groves R, et al. *Measurement Errors in Surveys*. Wiley: New York, 1991.
2. Bryant W, Ompad D, et al. Determinants of influenza vaccination in hard-to-reach urban populations. *Preventive Medicine*, 2006; **43**(1): 60-70.
3. Cochran W. *Sampling Techniques*. Chapters 5, 9, 10, 11. John Wiley and Sons. 1977
4. Cox B, Cohen S. *Mythological Issues for Health Care Surveys*. Marcel Drekker: New York, 1985.
5. Faugier J,Sargeant M. Sampling hard to reach populations. *Journal of Advanced Nursing*, 1997; **26**(4): 790-797.
6. Mines R, Mullenax N, Saca L. *The Binational Farmworker Health Survey*. California Institute for Rural Studies, Davis. 2001. http://www.cirsinc.org/Documents/Pub1001.2.pdf
7. Mines R. An Evaluation of the Gathering of Occupational Injury Data by the National Agricultural Workers Survey
8. Muhib F, Lin LS, Stueve A, Miller RL, Ford WL, Johnson WD, Smith PJ. A venue-based method of sampling hard-to-reach populations. *Public Health Reports* 2001; **116**(1) Suppl.: 216-222.
9. Owens L, Johnson TP, et al. (1999). Culture and Item Nonresponse in Health Surveys. *Seventh Conference on Health Survey Research Methods*, Hyattsville, Maryland, Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.
10. Spring M, Westermey J, et al. Sampling in difficult to access refugee and immigrant communities. *Journal of Nervous and Mental Illness* 2003; **191**(12):813-819.
11. Sukhatme PV. *Sampling Theory of Surveys, with Applications*. Iowa State College Press. Ames, Iowa; 1954
12. Thompson S. *Sampling*, Chapters 11, 12, 13. John Wiley and Sons. 1992
13. U.S. Department of Labor, Statistical Methods of the National Agricultural Workers Survey, http://www.doleta.gov/agworker/statmethods.cfm.
14. Villarejo D, Lighthall D, Williams D, Souter A, Mines R, Bade B, Samuels SJ, McCurdy SA. (2000) Suffering in Silence: A Report on the Health of California's Agricultural Workers. Davis, California Institute for Rural Studies.