C.   SAMPLING AND WEIGHTING

In this section we examine issues that were considered in developing the sample designs for the three Hispanic/Latino Adult Tobacco Survey (H/L ATS) case study sites. The important lesson is not how these issues were resolved in the three case studies, but how these issues relate to the population of interest. Most of these issues will be relevant in sampling other Hispanic and Latino target populations. It is recommended that a sampling statistician be consulted when the sampling plan for a specific survey is designed.

C.1   Sampling in Three Case Study Surveys

Three case studies are presented here that illustrate different approaches to developing probability samples of the Hispanic and Latino population. The three areas chosen for the case studies are (1) four boroughs of New York City; (2) Miami-Dade County, Florida; and (3) a compact group of three Hispanic neighborhoods called *colonias* in El Paso County, Texas, along the Texas-Mexico border. These locations were selected in part because they are typical of many such communities across the country; therefore, survey and sampling approaches that work in these three locations should work similarly in corresponding areas.

The main differences between the three surveys involve the mode of data collection and the recommended sampling frame. For determining the best mode of data collection for each area (telephone versus in-person), a crucial consideration is the percentage of the target area that is Hispanic. The New York case study represents a highly urban area with a slightly above-average density of Hispanic or Latino persons (29%). Miami, Florida, is likewise an urban area, but the density is notably higher (57%).

Both New York and Miami-Dade case studies target larger geographic areas with a smaller percentage of Hispanic persons than the *colonias*; therefore, a substantial number of households must be screened in these areas to locate Hispanic respondents. Telephone interviewing of a sample chosen from a standard list-assisted random-digit-dial (RDD) frame of telephone numbers is the choice for New York and Miami (Casady & Lepkowski, 1991) because in these sites (1) most households have telephone service and (2) the Hispanic and Latino population is relatively spread out. Telephone interviewing means there will be no interviewer travel costs and screening can be done efficiently.

By contrast, of the three case study areas, the *colonias* have the highest density of Hispanic and Latino persons (96%). Area sampling and face-to-face interviewing of selected residential dwellings is the choice for the *colonias* (Kish, 1965): many households there lack home telephones; the population resides in a small, contained area; and many persons there speak only Spanish. Face-to-face screening (area sampling) and in-person interviewing, therefore, should yield a better response rate than telephone interviewing. Moreover, this approach is relatively cost-efficient because sampling a compact area means interviewer travel cost will be low.

All survey research plans in all three sites share the following features:

- The sampling approach proposed for each site provides for a probability sample that can be considered representative of the target population.

- The sample frame of households developed in each site is random and representative.

- The target population for each survey is Hispanic/Latino residents aged 18 years or older and located by screening the households in the sample.

- The research objective of each survey is to profile patterns of adult tobacco use in the target population.

- The same survey materials are used in each site (with minor differences to accommodate the different modes of data collection).

- Targeted sample size in each location is 1,500 respondents, with one adult randomly selected from each sampled household.

- Respondents must speak either English or Spanish.

Table C-1.   Description of the Three H/L ATS Case Study Sites

Case study site	Approximate Hispanic adult pop. (year 2000)	Approximate Hispanic adult pop. (%)	Sampling frame(s)	Mode of data collection
New York City: boroughs of Brooklyn, Bronx, Manhattan, and Queens	1,227,200	29	List-assisted RDD	Telephone
Florida: the Miami portion of Dade County	971,800	57	List-assisted RDD	Telephone
El Paso County, Texas: *colonias* named Clint, San Elizario, and Socorro	23,500	96	List of U.S. Census blocks: lists residential dwellings in each sample block	In-person

C.1.1 New York, New York: Telephone Survey in Urban Area with Moderate Concentration of Hispanic and Latino Persons

New York's is a stratified simple random sample of enough telephone numbers to yield about 1,500 completed interviews with self-identified Hispanic residents aged 18 or older who can be reached by landline telephone in the four targeted boroughs of the Bronx, Brooklyn, Manhattan, and Queens.[2] For a site like New York, with its large area and low

---

[2] Limiting sampling to those households with telephone access creates some coverage bias in that it excludes Hispanic households without a home phone (Lessler & Kalsbeek, 1992). This source of bias can usually be controlled somewhat through weights calibration, by *poststratifying,* or *raking,* the weights as mentioned in Section C.3.3 (Kalton & Flores-Cervantes, 2003).

concentration of Hispanic and Latino persons, the topics that follow address increase of efficiency in the sampling approach and minimization of the costs of screening for Hispanic and Latino households.

Geographic Constraints

The New York case study targets the four New York boroughs with the highest density of Hispanic and Latino households. In this case, the researchers were satisfied that representative findings based on these four boroughs would meet their needs.

Use of List-assisted RDD

A list-assisted RDD telephone sample frame was recommended for New York. A list-assisted frame typically consists of those telephone numbers in telephone 100-banks[3] with at least one directory-listed telephone number (*list-assisted* because directory listings help identify the telephone prefixes to be sampled). List-assisted RDD sampling is recommended over other methods for several important reasons. List-assisted RDD sampling is more efficient than straight RDD sampling (choosing 10-digit phone numbers completely at random within the target area) because the list will contain a higher percentage of residential telephone numbers, and therefore less effort will be spent dialing nonproductive numbers. Sampling directly from a telephone directory would certainly result in more residential numbers, but it would exclude unlisted and unpublished phone numbers, a potentially serious source of bias (Kalsbeek & Agans, 2007). Similarly, Spanish surname lists drawn from published directories or other sources typically have limited coverage, which reduces the representativeness of the population. None of the three case studies recommends the use of surname lists.

Oversampling

Even after sampling is limited to these four boroughs, only 29% of the households contacted are expected to be Hispanic or Latino. A significant portion of the calling effort will have to be devoted to household screening. To improve these odds, it is possible to *oversample* Hispanic populations by identifying telephone prefixes known to contain higher concentrations of Hispanic households and sampling from these prefixes at a higher rate (Kalsbeek & Agans, 2007).

At the borough level, the percentage of Hispanic persons in the population for the Bronx (57%) is roughly twice that in the other boroughs (20%, 27%, and 26% for Brooklyn, Manhattan, and Queens, respectively). Oversampling phone numbers from the Bronx, therefore, may improve the calling efficiency for Hispanic households. To further increase the calling efficiency, oversampling by borough can be combined with oversampling of telephone prefixes known to correspond with higher concentrations of Hispanic households. These increases in calling efficiency come at a price, though, in terms of loss of precision

---

[3] A *100-bank* consists of those telephone numbers with the same first 8 digits of a 10-digit number.

(because of variable sampling probabilities and weights; Kalsbeek, 2003). The optimal allocation of sample between these methods also depends on the goals of the survey (e.g., whether separate estimates are sought for individual boroughs). Determining optimal sampling rates requires careful consideration of both statistical and practical implications. It is recommended that a sampling statistician and survey methodologist confer to discuss the pros and cons of any specific situation (Cochran, 1977).

Determining the Number of Selected Phone Numbers to Call

Although the telephone survey designs for New York and Florida target 1,500 completed interviews, the actual number of sample phone numbers that have to be called is much greater. The experience of prior telephone surveys with similar topics, target populations, or sample recruitment strategies can help with estimating the quantity of phone numbers that will be required. If *Y* is the expected ratio of number of respondents to number of assigned phone numbers, accounting for all sources of attrition combined, then to obtain 1,500 respondents one must assign 1,500/*Y* for calling in the site. When attrition patterns are likely to differ among the sampling strata that are used, one should separately estimate sample attrition and the number of selected phone numbers in each stratum or groups of strata where attrition is expected to be similar.

C.1.2 Miami, Florida: Telephone Survey in Urban Area with Higher Concentration of Hispanic and Latino Persons

Oversampling will result in some loss of precision; therefore, the value of oversampling areas with relatively high Hispanic concentrations must be balanced against the loss of precision due to variable weights (Kalsbeek, 2003). Because Miami has a greater concentration of Hispanic persons to begin with (57% as opposed to New York's 29%), a simpler sampling plan—just oversampling telephone prefixes with higher concentrations of Hispanic households—is recommended.

In both of these examples, the sole purpose of sample stratification is to facilitate an oversampling of Hispanic persons in the target area. Investigators may also be concerned about the precision of the estimate of tobacco use. If there is a large difference in the tobacco use levels between different parts of the target population, it may be of value to incorporate this information into the sampling plan. The merits of different sampling rates in a multistrata design would have to be evaluated by a sampling statistician in light of the specific characteristics of the target population. Suggested approaches for determining optimal sample allocations in different situations are provided in Section F.2: References and Resources.

C.1.3 El Paso, Texas: In-person Survey in Border Areas with High Concentration of Hispanic and Latino Persons

Multistage area sampling is commonly used to select households in face-to-face sample surveys, such as that for the El Paso site (see Kish, 1965). Area samples are most useful when the target area of the survey can be subdivided into a reasonably large number of well-defined geopolitical subunits for which population counts, maps, and other statistical data are available.

Two plausible alternatives to area sampling rely on different frame sources. One is sampling directly from postal mailing lists of residences (Iannacchione, Staab, & Redden, 2003), and the other is sampling parcels of land via electronic property tax files (Kalsbeek, Kavanagh, & Wu, 2004). Both of these alternatives have been shown to generate samples with very good coverage, to be simple and inexpensive to use, and to avoid the usually negative statistical effects of cluster sampling. Mailing lists have the added advantage of an easily accessible mailing address for sending advance letters, and the tax parcel approach has the added benefit of latitude-longitude coordinates to make sampled parcels easier to find.

Deciding on Sampling Units

Selection of an area sample of Hispanic persons in a local setting like the *colonias* typically calls for first choosing a sample of area subunits as primary sampling units (PSUs) and then randomly selecting a sample of residential dwellings as secondary sampling units (SSUs) in each selected PSU.[4] Each sample PSU is best chosen with a probability proportional to its size (i.e., a PPS, with *size* referring to the best measure of the number of Hispanic households in the PSU). An approximately equal number of Hispanic dwellings are then chosen within each PSU. The chosen dwellings come from a list frame separately and specially constructed by trained field staff who follow a rigorous protocol for list construction. The Census block is the most practical PSU for the H/L ATS in the El Paso site because (1) there are a sufficient number of them, (2) they are a tier of aggregation for urban sociodemographic characteristics from the decennial Census, and (3) there exist block maps with well-defined boundaries to facilitate sampling of dwellings within blocks.

Deciding on the Allocation Among Sampling Stages

A key feature of a multistage household sample is the allocation of the sample among stages. This allocation for the two-stage household sample design in the El Paso site is determined by the number of sample blocks (PSUs) and the average number of *selected* dwellings per sample PSU. These numbers are determined so that the total number of *responding* households will be 1,500. The experience of previously completed surveys can

---

[4] The terms *dwelling, housing unit, dwelling unit*, and *household* are synonymous, with the first three terms referring to the place where a group of related or unrelated individuals (comprising the household) resides.

help guide the decision about the number of selected dwellings to use as compared with the number of responding households required.

A good rule to follow is, the greater the number of sample PSUs one can afford, the better the statistical results from the sample will be. In practical terms, most good samples of this type strive for at least 50 sample PSUs and an average number of responding households per PSU no greater than 30.

Identifying Sampling Strata

Because the concentration of Hispanic persons is uniformly high in all three *colonias*, oversampling them by disproportionately sampling among *colonias* would not make household screening notably more efficient. However, PSU stratification by *colonia* would improve the precision of estimates of smoking prevalence for the population of Hispanic adults in the three *colonias* combined if there were substantial differences in smoking behavior among *colonias*.[5] The greater these differences, the greater the statistical benefit.

Stratification by other block-level characteristics available from the 2000 Census may also slightly improve the precision of H/L ATS estimates if those characteristics are correlated with smoking behavior measures of interest. Gender and other known predictors of smoking behavior that are available from Census block-level summary data could be used for this purpose.

Allocating Sample Size for Blocks Among Strata

Allocation of the sample of blocks among the PSU sampling strata will depend on which domains of the population are most important for analysis findings. If *colonias* and one or more other block-level characteristics are used to define strata, if the most important survey estimates are smoking prevalence rates for all Hispanic adults in the three *colonias* combined, and if the rates are not dramatically different among strata, then a proportionate allocation of the sample of blocks is the best choice. If, on the other hand, comparison of estimates among *colonias* is the highest priority, one third of the sample of blocks should be allocated to each of the *colonias,* and then the equal *colonia* sample sizes should be proportionately allocated among the strata within each *colonia*.

Selecting PPS Sample of Census Blocks as PSUs

An equal-probability sample of households, and its associated benefits, can be achieved within each stratum of a two-stage design (Kish, 1965). This outcome is accomplished by selection of blocks (PSUs) with PPS, with the best estimate of current household size as the size measure for PPS selection, and then selection of an equal number of dwellings within each selected block. A number of PPS selection methods could be used in this circumstance. One approach is PPS systematic sampling in which the PPS selection rule is applied to a

---

[5] Data from the 2000 Census indicates that for the *colonias* the percentage of the population that is Hispanic is 97.9% in San Elizario, 84.0% in Clint, and 96.4% in Socorro.

strategically ordered PSU frame by using a systematically selected sequence of numbers. Two alternatives are PPS with replacement sampling, in which it is possible to select a PSU multiple times, and PPS without replacement sampling, in which repeat selection is not allowed (Cochran, 1977). Each approach has its merits; these merits would have to be evaluated by a sampling statistician familiar with the specific target population.

Constructing a Sampling Frame for Second-stage Sampling

Choosing a subsample of dwellings may not be necessary in some sample blocks. When the average number of dwellings per block is small (e.g., fewer than 20), it may be more practical to include all dwellings in the SSU sample. The cutoff for identifying sample blocks not requiring subsampling depends on the targeted average number of responding households per sample PSU.

In those sample blocks where a subsample of dwellings is chosen, the frame for choosing dwellings may be constructed in a number of ways. The traditional approach has been to train field staff to list all dwellings by following a predetermined path around the boundary and internal streets of the block group. Although this approach produces a useful frame, it is relatively expensive to implement. Publicly available postal mailing lists and property tax parcel listings are alternatives.

Selecting Sample of Dwellings Within PSUs

Simple random sampling is typically applied to the block-specific frames just described. As with telephone sampling, the number of selected households in this final stage of household sampling must account for sample attrition due to ineligibility (e.g., vacant dwelling) and other reasons for nonresponse (e.g., refusal, not at home, unavailable) to result in 1,500 participating households.

C.2   Within-household Sampling

Households in H/L ATS samples are clusters of one or more Hispanic adults. One resident is randomly chosen for the survey interview in each household. Although there are several alternative methods for randomly choosing the resident, the H/L ATS screener employs the "*n*th-oldest adult" approach. This approach is relatively easy to use and is generally noninvasive, especially as compared with the household roster approach, though it can somewhat skew the sample.[6]

In its simplest version, the *n*th-oldest adult approach begins by determining the number of Hispanic adults residing in the household and then chooses a random number between one

---

[6] Some surveys request specific identifying information (e.g., the selected resident's first name or gender and age) to form a detailed household roster to use as the basis for resident selection. This is preferred from a technical standpoint to reduce gender bias, but asking for more clearly identifying information on a household roster in this way increasingly has been seen by respondents as prying or intrusive and has led to higher refusal rates. The H/L ATS screener does not use this method.

and the number of reported residents. The selected resident is designated by age, relative to the oldest resident. For example, if there are three eligible adults and the number 2 is randomly chosen, then the second-oldest adult is interviewed.

C.2.1 Reducing Gender Bias in Respondent Selection

The *n*th-oldest, next/last-birthday, and other respondent-selection methods that choose a resident at random often lead to a gender bias favoring females in the composition of the final respondent sample, if the gender of the selected resident is not provided. For example, populations with 50:50 splits between males and females can lead to 40:60 or even 30:70 splits in the respondent sample. One reason for this gender imbalance is that, all else being constant, females are more likely than males to be available for and respond to interview surveys. Another explanation for this gender imbalance is the tendency for the household resident completing the screener (more likely female than male) to claim to be the selected respondent if the selection method does not explicitly indicate who is to be chosen (Carr & Hertvik, 1993; Oldendick, Bishop, Sorenson, & Tuchfarber, 1988).

Gender bias can be reduced by more explicitly specifying who is selected. The H/L ATS screener asks for the number of adult Hispanic males and adult Hispanic females in the household. The interviewer can, for example, ask for the oldest female. With this approach, it is typical to require a separate random (i.e., Poisson) sampling decision for each household member, using selection probabilities that vary by subgroup characteristic (Lohr, 1999).

C.2.2 Respondent Selection in Multifamily Hispanic Households

In border areas like the *colonias,* there may be a higher frequency of multifamily households in the heavily Hispanic neighborhoods. Recently immigrated families tend to move in with more established residents, live with relatives, or "double up" with other recently immigrated families. A decision should be made early about whether the survey will recognize multiple families as separate sampling units or treat the sum of all adult residents as a single family for sampling purposes.

If the sum of adults is treated as a single family, the screener respondent must "count up" the total number of adults in residence, and then a single person is selected. Alternatively, multiple families at a single address may be considered separate reporting units for study data collection and therefore may be treated in effect as separate households. There are two options in this case: one is to conduct an interview with each family; the other is to first randomly choose one of the families and then select a respondent from among the residents of the selected family.

Selecting only one family avoids any estimate precision loss otherwise due to the clustering effect of interviewing multiple residents from the same household, but it can also contribute to reduced precision due to increased variation in selection probabilities among

respondents. Furthermore, selecting one family and one respondent avoids the practical difficulty of coding response dispositions from two respondents in the same household. Finally, allowing for multiple respondents per household makes it harder to predict how many interviews the sample will yield.

C.2.3 Additional Considerations

Two remaining points should be kept in mind. First, within-household sampling is another stage in the sample design. The probability of inclusion for any sample member in multistage designs is the product of selection probabilities for sample outcomes in each stage leading to the choosing of that member. The approach followed in selecting persons to interview is critical to determining the selection probabilities required to produce sample weights.

Second, in a computer-assisted telephone survey, the system will automatically choose whom to ask for, in accordance with the answers to screening questions. Operationalizing sampling procedures in an in-person screening, though, can be difficult. Interviewers must be provided a clear, easy-to-follow protocol for deciding what n is when they ask for the nth-oldest adult, man or woman.

C.3   Weighting Methods in the Three H/L ATS Case studies

During analysis, formulas are applied to sample data to produce estimates of the population characteristics. The statistical quality (or accuracy) of any survey estimate is measured by the size of its mean-squared error, which jointly depends on the precision (measured by variance or standard error of the estimate) and the bias of the estimate. Statistical inference based on probability samples offers an added advantage over inference using nonprobability samples: the analyst, using data from the chosen sample, can directly obtain measures of the statistical precision of estimates, although, like the survey estimates, these measures of precision are also estimates. These precision measures are required in order to produce confidence intervals, tests of hypothesis, and other statistical products of analysis. To supplement efforts called for by the survey design, the bias of survey estimates must be measured.

Appropriately estimating population characteristics and their precision requires that design features such as stratification, cluster sampling, and numerical measures of variable selection probabilities (i.e., leading to the computation of sample weights) be accommodated in analysis. Lohr (1999) offers a relatively recent review of the general design strategies and estimation issues related to sampling from finite populations. A more thorough discussion of other design issues in telephone surveys is given by Kalsbeek and Agans (2007). The representativeness of the selected sample may be altered by limitations in the selection and data-gathering processes, including frames that selectively cover the

target population, and differential nonresponse by members of the selected sample and among data items sought from responding sample members (Lessler & Kalsbeek, 1992).

C.3.1 Sample Weights

To produce representative findings, the analyst should (1) compute sampling weights to account for the process of sample selection and important composition-altering forces at work on the sample during the sampling and data collection processes, and (2) in analysis use statistical formulations that utilize these weights and appropriately account for stratification and cluster sampling in generating survey findings.

A sample weight is a statistical measurement linked to a data record for any survey respondent. In general terms, it is computed as the inverse of the adjusted probability of obtaining the data for the respondent. In most cases this probability is simply the respondent's original selection probability based on the sample design. The inverse probability, or base weight, is often adjusted to account for unintended sample imbalance arising during the conduction of the survey. More than one weight adjustment may be applied, and all are multiplicative.

Unless a weight is rescaled for analytic purposes (e.g., normalized to sum to the number of sample respondents), its value can be interpreted as an indication of the number of population members represented by the respondent. Separate sets of weights may be necessary when data are gathered for different types of data items associated with the respondent. For example, if data in a household survey are gathered for the selected households and for one resident chosen at random in each of those households, a separate set of weights is produced for the household data and the resident data.

C.3.2 Weight Calculation

Some combination of the following steps is typically followed to produce from a probability sample a set of weights for the "*i*th" individual-respondent data record, with the final adjusted weight being the product of the value generated in each step. If at all possible, all of the following steps should be completed on H/L ATS survey samples:

1. Base weight (determined by the probability of choosing the household and the method of respondent selection within the household).

2. Adjustment for nonresponse (to partially offset the biasing effects of differential response rates in the sample).

3. Adjustment for incomplete sample coverage (to partially correct for any bias due to differential coverage of the population by the list or lists from which the sample is chosen).

4. Adjustment to control variation among weights (to limit the loss in the precision of survey estimates due to widely variable sample weights).

5. Adjustment to calibrate the weights to the sampled population (to compensate for any sample imbalance not accommodated by the other adjustments).

Step 1 must always be completed in H/L ATS samples described in the case studies. For it to be completed, the sample design must qualify as a probability sample design, and steps followed in selecting the sample must be well documented so that selection probabilities can be determined for each survey respondent. Step 2 may be done if the sample can be subdivided into subgroups among which survey response rates differ. Step 3 will almost never be used for H/L ATS samples: computing it is practical only for sites where telephone sampling is done *and* for which there are external data on households with and without telephone access. Step 4 is particularly important in sites where the sample is significantly disproportionate (e.g., as a result of efforts to oversample Hispanic households). Step 5 is both important and difficult to implement for the typical target population of the H/L ATS.

C.3.3 Lack of Known Totals to Calibrate Weights

Step 5, sometimes referred to as *weighting up to known totals,* is a final correction that helps make the weighted data more representative of the target population. Weights calibration, however, requires high-quality external data on the target population distribution by population characteristics highly correlated with adult smoking behavior. Large, national-level population surveys commonly rely on information obtained from the most recent decennial Census, the Current Population Survey, or the American Community Survey. As the three case studies suggest, the H/L ATS is typically conducted at the substate, and often subcounty, level. It can be difficult to find a data source sufficiently current and of high quality to use in calibrating weights for a specific target population. Data from the most recent decennial Census are usually the best available option, although Census counts may not be altogether current.

Even if such data are available for a specific area, they may lack sufficient detail to correctly weight the data, as explained in the *Assessment of Major Federal Data Sets for Analyses of Hispanic and Asian or Pacific Islander Subgroups and Native Americans:*

> All of the major surveys use poststratification in the final stage of weighting to reduce sampling errors, and to compensate as much as possible for nonresponse and undercoverage. There are almost always separate poststratification cells for blacks, Hispanics, and all other race/ethnic groups. . . . The minority subgroups are almost always combined into categories like "total Hispanics" or "total other races. . . ." Subdomains such as Puerto-Ricans, Cuban-Americans, Central-Americans, etc., are thus combined into a single class, with identical weights. . . . If, in fact, some of these subgroups have lower response rates than the overall rate for the race/ethnic class, and are not separately adjusted, they will be underrepresented in the statistics. A similar situation exists with undercoverage. For example, if illegal aliens tend to avoid reporting (as seems likely) and if a higher proportion of Mexican-Americans are here illegally than in other Hispanic subpopulations (as is also likely), then the uniform weighting will slightly understate Mexican-Americans and overstate other Hispanic subgroups. (Waksberg, Levine, & Marker, 2000, sec. 2.7)

This statement is both an argument for achieving the highest response rates possible and a caveat about using known totals to weight the data.

C.3.4 Statistical Software for Complex Survey Designs

The sampling approaches described in the case studies are considered complex in that they may involve cluster selection, stratification, and sample weights. To prepare weights and weighted estimates from complex designs, one does best to use statistical software packages that rely on approximation or replication-based methods to estimate the variance of estimates (Wolter, 1985). A listing and several reviews of computer software that accommodates the sample design in this way are available online from the Survey Research Methods Section of the American Statistical Association at http://www.hcp.med.harvard.edu/statistics/survey-soft/.