CALIFORNIA HEALTH INTERVIEW SURVEY

CHIS 2003 METHODOLOGY SERIES

REPORT 1

SAMPLE DESIGN

October 2005

This report was prepared for the California Health Interview Survey by Ismael Flores Cervantes, R. Lee Harding and J. Michael Brick of Westat, Inc.

www.chis.ucla.edu

This report provides analysts with information about the sampling methods used for CHIS 2003, including both the household and person (within household) sampling. This report also provides a discussion on achieved sample size and how it compares to the planned sample size.

Suggested citation:

California Health Interview Survey. *CHIS 2003 Methodology Series: Report 1 - Sample Design*. Los Angeles, CA: UCLA Center for Health Policy Research, 2005.

The California Health Interview Survey is a collaborative project of the UCLA Center for Health Policy Research, the California Department of Health Services, and the Public Health Institute. Funding for CHIS 2003 came from multiple sources: the California Department of Health Services, the California Endowment, the National Cancer Institute, the Robert Wood Johnson Foundation, the California Office of the Patient Advocate, the U.S. Centers for Disease Control and Prevention, Kaiser Permanente, the Alameda County Health Care Agency, and L.A. Care Health Plan.

PREFACE

Sample Design is the first in a series of methodological reports describing the 2003 California Health Interview Survey (CHIS 2003). The other reports are listed below.

CHIS is a collaborative project of the University of California, Los Angeles (UCLA) Center for Health Policy Research, the California Department of Health Services, and the Public Health Institute. Westat was responsible for the data collection and the preparation of five methodological reports from the 2003 survey. The survey examines public health and health care access issues in California. The CHIS telephone survey is the largest state health survey ever undertaken in the United States. The plan is to monitor the health of Californians and examine changes over time by conducting periodic surveys in the future.

Methodological Reports

The first five methodological reports for CHIS 2003 are as follows:

- Report 1: Sample Design;
- Report 2: Data Collection Methods;
- Report 3: Data Processing Procedures;
- Report 4: Response Rates; and
- Report 5: Weighting and Variance Estimation.

The reports are interrelated and contain many references to each other. For ease of presentation, the references are simply labeled by the report numbers given above.

This report describes the procedures used to design and select the sample from CHIS 2003. An appropriate sample design is a feature of a successful survey, and CHIS 2003 presented many issues that had to be addressed at the design stage. This report explains why the design features of CHIS were selected and presents the alternatives that were considered.

The primary purpose of this report is to provide analysts information about the sampling methods used for CHIS 2003, including both the household and person (within household) sampling. In general terms, once a household was sampled, an adult within that household was sampled. If there were children and/or adolescents in the household, one child and/or one adolescent was eligible for sampling. This report also provides a discussion on achieved sample size and how it compares to the planned sample size.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

List of Tables

TABLE OF CONTENTS (continued)

List of Tables (continued)

List of Figures

1. CHIS 2003 DESIGN AND METHODOLOGY SUMMARY

1.1 Overview

The California Health Interview Survey (CHIS) is a population-based random-digit dial telephone survey of California's population that is conducted every two years. First conducted in 2001, CHIS is the largest health survey ever conducted in any state and one of the largest health surveys in the nation. CHIS is a collaborative project of the UCLA Center for Health Policy Research, the California Department of Health Services, and the Public Health Institute. CHIS collects extensive information for all age groups on health status, health conditions, health-related behaviors, health insurance coverage, access to health care services, and other health and development issues.

The CHIS sample is designed to provide population-based estimates for most California counties, all major ethnic groups, and several ethnic subgroups. The sample is designed to meet and optimize two goals: provide estimates for large- and medium-sized population counties in the state, and for groups of the smallest population counties; and provide statewide estimates for California's overall population, its major race/ethnic groups, as well as for several Asian ethnic groups. The resulting CHIS sample is representative of California's non-institutionalized population living in households.

This series of reports describes the methods used in collecting data for the 2003 California Health Interview Survey (CHIS 2003). CHIS 2001 is described in a series of methodology reports.[1] These reports describe the second CHIS data collection cycle, which was conducted between August 2003 and February 2004.

CHIS data and results are used extensively by many State agencies, local public health agencies and organizations, federal agencies, advocacy and community organizations and agencies, foundations, and researchers. They use these data in their own analyses and publications to assess public health and health care needs, to develop health policies, and to develop and advocate policies to meet those needs.

[1] California Health Interview Survey, CHIS 2001 Methodology Series: Report 1 - Sample Design, Report 2 – Data Collection Methods, Report 3 – Data Processing Procedures, Report 4 – Response Rates, and Report 5 – Weighting and Variance Estimation, Los Angeles, CA: UCLA Center for Health Policy Research, 2002.

1.2 Sample Design Objectives

The CHIS sample is designed to meet two objectives: (1) provide estimates for counties and groupings of counties with populations of 100,000 or more; and (2) provide estimates for California's overall population and its larger race/ethnic groups, as well as for several smaller ethnic groups. To achieve these objectives, CHIS relied on a multi-stage sample design. First, the state was divided into 41 geographic sampling strata, including 33 single-county strata and 8 groups that included the 25 other counties. Second, within each geographic stratum, households were selected through random-digit dial (RDD), and within each household, an adult (age 18 and over) respondent was randomly selected. In addition, in those households with adolescents (ages 12-17) and/or children (under age 12), one adolescent was randomly selected for interview and one child was randomly selected and the most knowledgeable parent of the child interviewed.

Table 1-1 shows the 41 sampling strata (i.e., counties and groups of counties that were identified in the sample design as domains for which separate estimates would be produced). A sufficient amount of sample was allocated to each of these domains to support the first sample design objective. These strata were also used for the CHIS 2001 sample; because of funding limitations, the sample sizes allocated to most strata for CHIS 2003 were smaller than in 2001.

Table 1-1. California county and county group strata used in the CHIS 2003 sample design

1. Los Angeles
2. San Diego
3. Orange
4. Santa Clara
5. San Bernardino
6. Riverside
7. Alameda
8. Sacramento
9. Contra Costa
10. Fresno
11. San Francisco
12. Ventura
13. San Mateo
14. Kern
15. San Joaquin
16. Sonoma
17. Stanislaus
18. Santa Barbara
19. Solano
20. Tulare
21. Santa Cruz
22. Marin
23. San Luis Obispo
24. Placer
25. Merced
26. Butte
27. Shasta
28. Yolo
29. El Dorado
30. Imperial
31. Napa
32. Kings
33. Madera
34. Monterey, San Benito
35. Del Norte, Humboldt
36. Lassen, Modoc, Siskiyou, Trinity
37. Lake, Mendocino
38. Colusa, Glen, Tehama
39. Sutter, Yuba
40. Plumas, Nevada, Sierra
41. Alpine, Amador, Calaveras, Inyo, Mariposa, Mono, Tuolumne

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

The samples in Los Angeles and Alameda Counties were enhanced with additional funding to allow sub-county geographic estimates, in Los Angeles at the Service Planning Area (SPA) level and in Alameda for the cities of Oakland and Hayward as well as the remainder of the county. These samples were implemented with and incorporated into the original statewide RDD sample.

To accomplish the second objective, larger sample sizes were allocated to the more urban counties where a significant portion of the state's Latino, African American and Asian ethnic populations reside. To increase the precision of the estimates for Koreans and Vietnamese, areas with relatively high concentrations of these groups were sampled at higher rates; these geographic samples were supplemented by phone numbers for group-specific surnames drawn from listed telephone directories to increase the sample size and precision of the estimates for these two groups.

1.3 Data Collection

To capture the rich diversity of the California population, interviews were conducted in five languages: English, Spanish, Chinese (Mandarin and Cantonese dialects), Vietnamese, and Korean. These languages were chosen based on research that identified the languages that would cover the largest number of Californians in the CHIS sample that either did not speak English or did not speak English well enough to otherwise participate.

Westat, a private firm that specializes in statistical research and large-scale sample surveys, conducted the CHIS 2003 data collection. Westat staff interviewed one randomly selected adult in each sampled household. In those households with children (under age 12) or adolescents (ages 12-17) associated with the sampled adult[2], one child and one adolescent were randomly sampled, so up to three interviews could have been completed in each sampled household. The sampled adult was interviewed, and the parent or guardian most knowledgeable about the health and care of the sampled child was interviewed. The sampled adolescent responded for him or herself, but only after a parent or guardian gave permission for the interview. Table 1-2 shows the number of completed adult, child, and adolescent interviews in CHIS 2003, by the type of sample (RDD or supplemental sample).

[2] Only children for whom the sampled adult was parent or legal guardian were sampled. The CHIS 2003 sample weights account for this sampling procedure.

Table 1-2. Number of completed CHIS 2003 interviews by type of sample, instrument

Type of sample	Adult	Child	Adolescent
Total RDD + supplemental cases	42,044	8,526	4,010
RDD	41,818	8,480	3,996
Supplemental samples:			
Korean	112	24	6
Vietnamese	114	22	8

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

Interviews done in all languages were administered using Westat's computer-assisted telephone interviewing (CATI) system. The average adult interview took 33 minutes to complete. The average child and adolescent interviews took 14 minutes and 21 minutes, respectively. Interviews in the non-English languages generally took longer to complete. Approximately 11 percent of the adult interviews were completed in a language other than English, as were 21 percent of all child (parent proxy) interviews and 7 percent of all adolescent interviews.

Table 1-3 shows the major topic areas for each of the three survey instruments (adult, child, and adolescent).

1.4 Response Rate

The overall response rate for CHIS 2003 is a composite of the screener completion rate (i.e., success in introducing the survey to a household and randomly selecting an adult to be interviewed), and the extended interview completion rate (i.e., success in getting the selected person to complete the full interview). To maximize the response rate, especially at the screener stage, an advance letter (in five languages) was mailed to all sampled telephone numbers for which an address could be obtained from reverse directory services. An advance letter was mailed for approximately 72 percent of the sampled telephone numbers. In 2003, the screener completion rate was 55.9 percent[3], and the rate was higher for those households that could be sent the advance letter. The extended interview completion rate was 60.0 percent for the adult survey. Multiplying the screener and extended rates gives an overall response rate of 33.5 percent. Response rates vary by sampling stratum.

[3] In CHIS 2003, households that refused at the screener level were subsampled and only the subsampled households were called again in an attempt to convert them to respondents. The response rates are weighted to account for this subsampling.

Table 1-3. CHIS 2003 Survey topic areas by instrument

HEALTH STATUS	**ADULT**	**TEEN**	**CHILD**
General health status, height and weight	✓	✓	✓
Emotional health		✓	
Days missed from school due to health problems		✓	✓
HEALTH CONDITIONS	**ADULT**	**TEEN**	**CHILD**
Asthma	✓	✓	✓
Heart disease, high blood pressure, epilepsy	✓		
Diabetes	✓	✓	
Physical disability/need for special equipment	✓	✓	✓
Elder health (stroke, falls, incontinence)	✓		
Parental concerns with child development, attention deficit disorder (ADD)			✓
HEALTH BEHAVIORS	**ADULT**	**TEEN**	**CHILD**
Dietary intake		✓	✓
Physical activity and exercise		✓	✓
Walking for transportation and leisure	✓		
File and pneumonia immunization	✓		
Alcohol and tobacco use	✓	✓	
Drug use		✓	
Sexual behavior, STD testing, birth control practices	✓	✓	
WOMEN'S HEALTH	**ADULT**	**TEEN**	**CHILD**
Pap test screening, mammography screening, self-breast exam	✓		
Emergency contraception, pregnancy status	✓	✓	
Menopause, hormone replacement therapy (HRT)	✓		
CANCER HISTORY AND PREVENTION	**ADULT**	**TEEN**	**CHILD**
Cancer history of respondent	✓		
Colon cancer screening, prostrate cancer (PSA) test	✓		
DENTAL HEALTH	**ADULT**	**TEEN**	**CHILD**
Last dental visit, could not afford care, missed school/work days	✓	✓	✓
Dental insurance coverage	✓	✓	✓
INJURY/VIOLENCE	**ADULT**	**TEEN**	**CHILD**
Serious injuries (frequency, cause)		✓	✓
Injury prevention behaviors (bike helmets, seatbelts)		✓	✓
Infant-toddler home safety			✓
Interpersonal violence		✓	

Table 1-3. (Continued)

ACCESS TO AND USE OF HEALTH CARE	**ADULT**	**TEEN**	**CHILD**
Usual source of care, visits to medical doctor	✓	✓	✓
Emergency room visits	✓	✓	✓
Delays in getting care (prescriptions, tests, treatment)	✓	✓	✓
Health care discrimination due to race or ethnic group	✓		
Communication problems with doctor	✓	✓	✓
Ability and parental knowledge of teen contacting a doctor		✓	
Child immunization reminders			✓
HEALTH INSURANCE	**ADULT**	**TEEN**	**CHILD**
Current insurance coverage, spouse’s coverage, who pays for it	✓	✓	✓
Health plan enrollment, characteristics and assessment of plan	✓	✓	✓
Whether employer offers coverage, respondent/spouse eligibility	✓		
Coverage over past 12 months	✓	✓	✓
Reasons for lack of insurance	✓	✓	✓
EMPLOYMENT	**ADULT**	**TEEN**	**CHILD**
Employment status, spouse’s employment status	✓		
Work in last week, industry and occupation	✓		
Hours worked at all jobs	✓	✓	
INCOME	**ADULT**	**TEEN**	**CHILD**
Respondent and spouse’s earnings last month before taxes	✓		
Household income (annual before taxes)	✓		
Number of persons supported by household income	✓		
Assets	✓		
PUBLIC PROGRAM ELIGIBILITY	**ADULT**	**TEEN**	**CHILD**
Household poverty level (100%, 130%, 200%, 300% FPL)	✓		
Program participation (TANF, CalWorks, Public Housing, Food Stamps, SSI, SSDI, WIC)	✓	✓	✓
Assets, alimony/child support/social security/pension	✓		
Reason for Medi-Cal non-participation among potential eligibles	✓	✓	✓
FOOD INSECURITY/HUNGER	**ADULT**	**TEEN**	**CHILD**
Availability of food in household over past 12 months	✓		
PARENTAL INVOLVEMENT	**ADULT**	**TEEN**	**CHILD**
Parental presence after school, parental knowledge of whereabouts and activities		✓	
Child’s activities with family			✓
NEIGHBORHOOD AND HOUSING	**ADULT**	**TEEN**	**CHILD**
Neighborhood cohesion	✓		
Neighborhood safety	✓	✓	
Neighborhood characteristics for children			✓
Length of time at current address/neighborhood, type of housing	✓		
Home ownership, number of rooms, amount of mortgage/rent	✓		

Table 1-3. (Continued)

CHILD CARE	ADULT	TEEN	CHILD
Current child care arrangements			✓
Child care over past 12 months			✓
Reason for lack of childcare			✓
RESPONDENT CHARACTERISTICS	**ADULT**	**TEEN**	**CHILD**
Age, gender, height, weight, education	✓	✓	✓
Race and ethnicity	✓	✓	✓
Marital status	✓		
Sexual orientation	✓		
Citizenship, immigration status, country of birth, English language proficiency	✓	✓	✓

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

The CHIS response rate is comparable to response rates of other scientific telephone surveys in California, such as the California Behavioral Risk Factor Surveillance System (BRFSS) survey. California as a whole, and the state's urban areas in particular, are among the most difficult parts of the nation in which to conduct telephone interviews. Survey response rates tend to be lower in California than nationally, and over the past decade response rates have been declining both nationally and in California.

One way to judge the representativeness of a population survey is to "benchmark" its results against those of other reliable data sources. The CHIS 2001 sample yielded unweighted and weighted population distributions and rates that are comparable to those obtained from other sources. The demographic characteristics of the CHIS 2001 sample (such as race, ethnicity, and income) are very similar to those obtained from 2000 Census data. CHIS 2001 respondents also have health characteristics and behaviors that also are very similar to those found in other reliable surveys, such as the California BRFSS. An extensive benchmarking project is being undertaken for the 2003 California Health Interview Survey.

Adults who had completed at least 80 percent of the questionnaire (i.e., through Section I on health insurance) after all followup attempts were exhausted to complete the full questionnaire were counted as "complete." At least some items in the employment and income series or public program eligibility and food insecurity series are missing from these cases.

Proxy interviews were allowed for frail and ill persons over the age of 65 to avoid biases for health estimates for elderly persons that might otherwise result. Eligible selected persons were recontacted and offered a proxy option. For 171 elderly adults, a proxy interview was completed by either a spouse/partner or adult child. Only a subset of questions identified as appropriate for a proxy respondent were administered. (Note: The questions not administered are identified in their response set as being skipped (denoted by a value of “-2”) because a proxy is responding for the selected person.)

1.5 Weighting the Sample

To produce population estimates for the RDD CHIS results, weights are applied to the sample data to compensate for a variety of factors, some directly resulting from the design and administration of the survey. The sample is weighted to represent the non-institutionalized population for each sampling stratum and statewide. Sample weighting was carried out in CHIS 2003 to accomplish the following objectives:

- Compensate for differential probabilities of selection for households and persons (Note: telephone numbers for which addresses could be found and advance letters mailed were assigned a higher probability of selection than those without addresses);
- Reduce biases occurring because nonrespondents may have different characteristics than respondents;
- Adjust, to the extent possible, for undercoverage in the sampling frames and in the conduct of the survey; and
- Reduce the variance of the estimates by using auxiliary information.

As part of the weighting process, a household weight was created for all households that completed the screener interview. This household weight is the product of the “base weight” or the inverse of the probability of selection of the telephone number and adjustment factors computed for the following weight adjustments:

- Subsampling for numbers with addresses;
- Multiple chances of being selected in the RDD and supplemental samples;
- Unknown residential status;
- Subsampling screener refusals for conversion attempt;
- Screener interview nonresponse;
- Multiple telephone numbers; and
- Household poststratification.

The resulting poststratified household weight was used to compute a person-level weight. This person-level weight includes weight adjustments for the within-household sampling of persons and nonresponse. The final step is to adjust the person-level weight using a raking method so that the CHIS estimates are consistent with population control totals. Raking is an iterative procedure that forces the CHIS weights to sum to known totals from auxiliary data sources. The procedure requires iteration to make sure all the control totals or dimensions of raking are simultaneously satisfied (within a specified tolerance).

The control totals or raking dimensions used in CHIS 2003 were created primarily from the 2003 California Department of Finance estimates of the numbers of persons by age, race, and sex, and from the 2000 Census of Population counts from the U.S. Census Bureau. The 14 dimensions are combinations of demographic variables (age, sex, race, and ethnicity), geographic variables (county, city, and, in Los Angeles County, Service Planning Area), household composition (presence of children and adolescents in the household), and socio-economic variables (home ownership and education). The socio-economic variables are included to reduce biases associated with excluding households without a telephone number from the survey. One of the limitations of using the Department of Finance data is that it includes about 2.4 percent of the population of California who live in “group quarters” (i.e., persons living with 9 or more unrelated persons). These persons were excluded from the CHIS sample and, as a result, the number of persons living in group quarters had to be estimated and removed from the control totals prior to raking.

1.6 Imputation Methods

To enhance the utility of the CHIS 2003 data files, missing values were replaced through imputation for nearly every variable. This was a massive task designed to eliminate missing values in all source variables. Westat imputed values for variables used in the weighting process, and the UCLA staff imputed values where missing due to item nonresponse for nearly all other variables.

Two different imputation procedures were used by Westat prior to delivering the data to UCLA to fill in missing responses for items in CHIS 2003 that were essential for weighting the data. The first imputation technique is a completely random selection from the observed distribution of the respondents. This method is used only for a few items when the percentage of the items that are missing is very small. For example, when imputing the missing values for self-reported age which had a very low item non-response rate, the distributions of the responses for age by type of interview (adult, child, or adolescent) were used to randomly assign an age using probabilities associated with these distributions.

The second technique is hot deck imputation without replacement. The hot deck approach is probably the most commonly used method for assigning values for missing responses in large-scale household surveys. With a hot deck, a value reported by a respondent for a particular item is assigned or donated to a “similar” person who did not respond to that item. The characteristics defining “similar” vary for different variables. To carry out hot deck imputation, the respondents to an item form a pool of donors, while the nonrespondents are a group of recipients. A recipient is matched to the subset pool of donors based on household and individual characteristics. A value for the recipient is then randomly imputed from one of the donors in the pool. Once a donor is used, it is removed from the pool of donors for that variable. Hot deck imputation was used to impute race, ethnicity, home ownership, and education in CHIS 2003.

The UCLA staff imputed missing values through a hierarchical sequential hot deck method with donor replacement. This method rank-orders the control variables from the most essential to the least essential, allowing the control variables to be dropped if the imputation conditions (such as minimal number of donors or no missingness in control variables) are not met in the imputation process. The control variables are dropped one at a time sequentially, starting from the least essential. CHIS incorporated an automated data quality control check both before and after the imputation process.

Imputation flags for CHIS source variables are included in separate data files to identify all imputed values.

1.7 Methodology Report Series

A series of five methodology reports are available with more detail about the methods used in CHIS 2003:

- Report 1 – Sample Design;
- Report 2 – Data Collection Methods;
- Report 3 – Data Processing Procedures;
- Report 4 – Response Rates; and
- Report 5 – Weighting and Variance Estimation.

For further information on CHIS data and the methods used in the survey, visit the California Health Interview Survey Web site at www.CHIS.ucla.edu or contact CHIS at CHIS@ucla.edu.

2. TELEPHONE SAMPLING METHODS

This chapter describes the two general sampling methods used in the CHIS 2003 telephone survey. CHIS 2003 consisted of a telephone random digit dialing (RDD) sample[4] combined with Korean and Vietnamese surname list samples. The RDD sample was drawn using a list-assisted RDD approach, whereas the list samples were drawn from separate surname lists of telephone numbers. The first section below describes the list-assisted RDD sampling and the procedures implemented in CHIS 2003 to save costs by reducing the number of calls to ineligible telephone numbers in this sample. The second section reviews the sampling alternatives that were considered for supplementing the RDD sample to increase the sample size for Koreans and Vietnamese. This section also gives the rationale for deciding on the approach used for the supplemental samples.

Households without a telephone were not sampled for CHIS 2003, which could give rise to bias in the estimates. The bias is related to the percentage of households without telephones and the difference in characteristics of the telephone and nontelephone households. In the 2000 Census approximately 1.5 percent of households in California are without telephones. Recent evidence (Ford 1998; Anderson, Nelson, and Wilson 1998) shows that the health characteristics of those with and without telephones are not as different as they had been in the past. Based on these factors, it is unlikely that most estimates from CHIS will have substantial bias because nontelephone households are not sampled. However, some estimates that are very directly correlated to income may be subject to greater biases due to this form of undercoverage. To mitigate the effects of excluding households without telephones, special weighting procedures were used and these are described in *CHIS 2003 Methodology Series: Report 5 – Weighting and Variance Estimation*.

Another source of bias is related to the increased popularity of cellular telephones. A new group of households is not covered in traditional RDD surveys because the samples are selected from telephone numbers with exchanges assigned to landline telephones. As a larger proportion of households have cell phones only, the undercovered group becomes more heterogeneous and the sample becomes more difficult to adjust for undercoverage. Blumberg et al. (2004) presents the most relevant data with respect to the cell-only population in 2003. They show that about 3.2 percent of households nationally had cell phones only in the first six months of 2003. Tucker et al. (2004) provide details on the percentages of households with different types of telephone service and the characteristics of those with

[4] Supplemental samples selected by taking larger samples in geographic areas are considered part of the RDD Sample

cell phones only, but for 2004.[5] Both papers point out that the cell-phone-only households do have different characteristics from those with landlines, and the Blumberg et al. paper shows that some health characteristics such as health insurance coverage are different for this group of households. At this time, no special weighting adjustments have been developed to address this coverage problem.

In many household surveys, persons who do not speak English and in some cases who do not speak Spanish are sampled but never interviewed because of language difficulties. While technically we prefer to treat this as a nonresponse problem (language problem cases are considered nonrespondents), it could easily be thought of as a coverage problem since none of the persons with language difficulties are interviewed. In CHIS 2003, significant efforts were expended to limit this source of bias by interviewing in multiple languages. This effort should eliminate a large source of the bias that might result from conducting interviews in English or English and Spanish only.

2.1 List-Assisted Random-Digit-Dial Sampling

List-assisted sampling is a procedure for RDD telephone surveys made possible by recent technological developments (Casady and Lepkowski, 1993). In list-assisted sampling, the set of all telephone numbers in operating telephone prefixes is considered as composed of 100-banks. Each 100-bank contains the 100 telephone numbers with the same first eight digits (i.e., the identical area code, telephone prefix, and first two of the last four digits of the telephone number). All 100-banks with at least one residential number listed in a published telephone directory are identified. The sampling frame is restricted to these 100-banks. A simple random or a systematic sample of telephone numbers is selected from this frame.

List-assisted RDD sampling is currently the standard method of choice for telephone surveys. It results in an unclustered sample that can be released to interviewers once the sample of telephone numbers is chosen. These are both important features not shared by the Mitofsky-Waksberg method that used to be the standard RDD sampling technique (Brick and Waksberg, 1991). Furthermore, the working residential rate among sampled numbers (critically important in determining the cost of an RDD sample) is comparable to the Mitofsky-Waksberg technique. The only disadvantage is a small amount of undercoverage because telephone numbers in 100-banks with no listed telephone numbers are not sampled. Studies have been carried out on the potential losses associated with this truncated form of list-assisted sampling (Brick, et al., 1995; Giesbrecht, et al., 1996). The studies show only about two to

[5] The February 2004 Current Population Survey estimates approximately that 6% of households in the US have cell only telephone service.

four percent of households is excluded by this method. Furthermore, the households that are excluded do not appear to be very different from those included in the frame. As a result, the bias due to this method of sampling is considered negligible for most estimates.

When using a list-assisted approach, special procedures can be implemented to reduce costs before data collection to reduce costs. Some nonresidential telephone numbers can be "purged" or excluded from dialing prior to the start of data collection increasing the efficiency of contact efforts. The procedure used in CHIS 2003, called Genesys ID Plus, is offered by Market Systems Group[6] (MSG), who also provided the sampling frames. The ID Plus process is an enhancement to the Genesys ID process used in CHIS 2001. ID Plus takes advantage of recent developments in linking of data sources and technology to provide more result codes than the previous process and to classify a larger proportion of numbers as nonproductive (i.e., business and nonworking numbers). The components of ID Plus are White and Yellow Pages matches and tritone tests. With the White and Yellow Pages matches, a telephone number is considered a nonresidential business number if it is listed in a Yellow Pages directory but not in a White Pages directory. Numbers so designated are not dialed during data collection. The tritone test dials each telephone number that is not listed in either the White or Yellow Pages and allows up to two rings. Any telephone number where a tritone (the distinctive three-bell sound heard when dialing a nonworking number) is encountered in two separate tests is considered nonworking, and is not dialed during data collection. During the tritone test, an MSG representative picks up if the telephone call is answered and attempts to ascertain whether the telephone number is business or residential. Table 2-1 shows the ID Plus result codes as well as the distribution of the sampled telephone numbers. In CHIS 2003 a total of 38.49 percent of the sampled numbers (result codes LB, UB, FM, NR and NW) were excluded from dialing.

Table 2-1. ID Plus result codes and their distribution for the CHIS 2003 RDD sample

ID Plus result code	Description	Percentage
LR	Listed residential	27.31
LB	Listed business	6.12
UR	Unlisted residential	10.02
UB	Unlisted business	5.39
FM	Fax/modem	3.42
LA	Language barrier	0.52
NR	No ring back	0.25
NW	Nonworking	23.31
DK	Undetermined: No answer/busy	22.68
PM	Privacy manager	0.97
Total	--	100.00

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

. [6] Market Systems Group (MSG)/Genesys Sampling Systems provides a wide variety of services to the survey research community. Among these services, MSG maintains databases for sample selection in telephone surveys.

Another procedure used in the CHIS 2003 RDD sample, as in CHIS 2001, involves subsampling from the numbers not purged by ID Plus. First, each number is classified by whether a mailing address[7] can be associated with it, i.e., whether it is "mailable." Telephone numbers were subsampled at differential rates depending on whether they were mailable. Since mailable telephone numbers are much more likely to be residential, all of these telephone numbers were retained in the sample. The telephone numbers without mailable addresses are less likely to be residential so the cost of finding a residential number is greater in this substratum. For the CHIS 2003 RDD sample, 75 percent of the non-mailable addresses were retained in the subsample.

Another technique to improve sample efficiency in CHIS 2003 is subsampling of refusals[8]. In this procedure, not used in CHIS 2001, a larger sample of telephone numbers than would otherwise be selected is drawn in the first phase. Each number in this first-phase sample is randomly assigned to the second-phase subsample or not. When refusals are encountered at the screening stage of data collection, only numbers in the subsample are eligible for refusal conversion follow-up. The numbers subsampled for refusal follow-up are fielded first so that any refusal cases can be worked completely. The principles for refusal subsampling are well-established (e.g., Hansen and Hurwitz 1946; Elliott, Little, and Lewitzky 2000) and the method is used in other surveys, including the American Community Survey conducted by the U.S. Census Bureau.

The rationale for refusal subsampling depends on two observations: refusal cases comprise the majority of screener nonresponse in CHIS; and substantial effort is expended to gain cooperation in households in which a member refuses to participate in the study at the screener level. The subsampling of refusals shifts some resources from the less productive, labor-intensive task of refusal conversion to the more productive task of completing extended interviews. A weighting adjustment accounts for the subsampling, so that those cases that refuse and are subsampled are weighted to represent themselves and the cases that refuse and are not subsampled. This weighting decreases the precision of the survey estimates, but only very slightly. The weighting method is discussed in *CHIS 2003 Methodology Series: Report 5 – Weighting and Variance Estimation*. A subsampling rate of approximately 60 percent was used in CHIS 2003, meaning that 60 percent of the refusal cases at the screener level were eligible for refusal conversion efforts. This rate is expected to result in increases in the standard error of the estimates of less than a 3 percent.

[7] Several companies provide services of this type in which a telephone number is matched to commercially-available files of addresses.

[8] Refusal conversion subsampling and refusal conversion were used only in the RDD sample and not in the surname list samples in CHIS 2003.

The RDD sample also includes additional samples in geographic areas to meet special needs. The areas that had larger RDD samples for this purpose were the Special Planning Area (SPA)[9] of Antelope Valley in Los Angeles County and the cities of Hayward and Oakland in Alameda County. Since the sampling procedures for these geographic supplemental samples simply involved taking larger samples, the methods discussed above for the RDD sample also apply to these areas.

2.2 Supplemental Sampling

An important goal of CHIS 2003 was to produce reliable estimates for Koreans and Vietnamese in California with a goal of 500 adult interviews for each group (see Table 1-2). These two ethnic groups are important for analytic reasons, but constitute a small proportion of the total California population. The expected sample yield from the overall CHIS 2003 RDD sample was too small to support making inferences for these subgroups at the desired level of precision, so sampling methods for rare populations were considered for increasing the yield. Kalton and Anderson (1986) and Sudman, Sirken, and Cowan (1988) are general references for sampling rare populations.

The within-county goals of CHIS 2003 included a supplemental sample of 100 completed interviews with African-American adults in the city of Hayward. The methods for sampling rare populations were also considered for this group.

Several sampling strategies were considered to increase the sample yield for the race-ethnic samples in CHIS 2003. Two strategies adopted for the Korean and Vietnamese sample were stratified disproportional sampling and multiple frame sampling. Stratified disproportional sampling in combination with household screening was used in the Hayward African-American supplemental sample. Among the strategies considered but not adopted for these groups were multiplicity or network sampling and snowball sampling. Each of these strategies considered is described below, along with the reasons for choosing those that were adopted.

The screening strategy increases the initial survey sample size to support the smallest or rarest target subgroup. Once a sampled telephone number is determined to be residential, screening questions classify the household according to the presence of one or more adults in the target population(s). If the household contains a member of the rarest subgroup, it is retained. Otherwise, it is

[9] Special Planning areas (SPAs) are basic geographical organization units for planning and evaluation on public health issues in Los Angeles County. Each SPA is made up of one or more of the 24 Health Districts that are maintained as units for data collection, analysis and delivery of core public health services.

subsampled; the subsampling rates may be different for different target groups. This strategy is relatively simple to implement and has good statistical properties, except that measurement error may be introduced by asking a question about race or ethnicity in the beginning of a telephone interview. Because the data collection costs for this strategy increase dramatically the rarer the target population, it was not explored extensively for the Vietnamese and Korean samples in CHIS 2003. It was used, however, for African-Americans in Hayward, both because they were not as rare a population and because there was no surname list comparable to those used for the Korean and Vietnamese samples for a multiple frame approach.

Another sampling strategy considered but not adopted is called multiplicity or network sampling. In this approach, each sampled adult identified as being in the target group would be asked to identify other individuals in the target group and not living in his/her household. These individuals would then be contacted. This method is an inexpensive way of locating and interviewing a larger sample. The identification of other members of the target group is a key part of the sample selection process. Linkages to the other individuals must be unambiguously defined to compute unbiased estimates in accordance with the requirements of probability sampling. For example, most adults in California can only be sampled once in the RDD sample through the household's telephone number. Using multiplicity sampling, an adult in the target group could be sampled not only by selecting his/her household's own telephone number, but also as a result of linking that number to other adults in California. The links in multiplicity sampling are usually immediate relatives. Most often, sampled individuals are asked about their children, parents, or siblings not living in their own household and these individuals constitute a network. The probability of each individual in this network (including the adult sampled from the original telephone number) is then computed using the reported size of the network. An alternative version of multiplicity sampling is sometimes called "snowball" sampling, which is similar but does not attempt to define the size of the networks. Hence, it results in a nonprobability sample, which was not considered acceptable for CHIS.

A number of obstacles made multiplicity sampling unappealing for CHIS. The most serious impediment would be nonresponse in several manifestations. The first and most obvious issue is the willingness of race-ethnic groups to identify all their relatives who live in California and provide enough information to the interviewers so that they can be contacted. A related issue is the willingness of the linked siblings to respond to the interview. Recently Institutional Review Boards have also raised some confidentiality and privacy concerns regarding this method of sampling that requires one person to nominate another person to be a respondent to a survey especially on sensitive topics. Finally, the costs and yields for this approach could not be accurately estimated in advance because network sampling is untested in the CHIS setting.

One of the sampling strategies adopted in CHIS 2003 for the Korean and Vietnamese supplemental samples was a stratification approach, also known as disproportionate sampling. Under this scheme, auxiliary information is used to classify telephone exchanges (or banks of telephone numbers) by the proportion of members of the target groups residing in these exchanges. After classifying the exchanges into strata, the telephone numbers in the exchanges with a relatively high proportion of members are sampled at a higher rate than the numbers in the other strata. If the data used to stratify the numbers are accurate, then the telephone numbers in the oversampled exchanges are more likely to result in interviews with members of the target groups.

Stratified disproportionate sampling was the method used to enhance the yield of the supplemental Korean and Vietnamese sample and for the Hayward African-American supplemental sample in CHIS 2003. Information from the Census 2000 was used to stratify telephone exchanges into high and low concentrations of Koreans and/or Vietnamese in four counties[10], and of African-Americans in Hayward. These data were not available in 2001 and thus this method could not be used in CHIS 2001. Details of the implementation in CHIS 2003 are presented in section 3.3. Because only four counties in California had areas with sufficient concentrations of Koreans and Vietnamese to warrant stratification, and because this approach was not used in CHIS 2001, we recommended combining it with another approach for enhancing the Korean and Vietnamese sample yield.

The other approach used for the Korean and Vietnamese samples is based on the concept of a dual frame design. Under this design, one sampling frame, in this case that used for the CHIS 2003 RDD sample, is supplemented with a much less expensive sample from a list of telephone numbers likely to include members of the target group(s). The list frame does not have to be complete to be useful, although the more complete the list the greater the potential for increasing the precision of the estimates. The composition of the list affects its efficiency (that is, the proportion of sampled numbers that lead to a member of the target group), but not the ability to produce unbiased estimates. Unbiased estimates can be produced if the list membership of every sampled unit (telephone number) from the other (RDD) frame can be determined. Of course, if the list only contains members of one subgroup of the target group, the efficiency for many types of analysis may be adversely affected. In most applications, the cost of data collection using a list is dramatically lower than the cost for screening for members of the rare population. See *CHIS 2001 Methodology Series: Report 2 – Data Collection* for a comparison of per-completed-interview costs for the RDD, Korean, and Vietnamese surname list samples.

[10] Stratified disproportionate sampling was used in four counties that cover more that 75 percent of the Korean and Vietnamese population in California. These counties were Los Angeles, San Diego, Orange, and Santa Clara.

In a dual frame approach, the characteristics of the list are very important and worth reviewing in some detail. The first characteristic is that the list must contain the telephone number for members of the target group so the sample from the list can be interviewed. The telephone numbers are also needed for estimation purposes, as described in *CHIS 2003 Methodology Series: Report 5 – Weighting and Variance Estimation*. A second important property of the list is the proportion of the population of interest it contains. Lists that are more complete make the sampling process more efficient. A third property of the list is the need to cover a relatively broad spectrum of types of the target group members. Finally, the accuracy of the lists in identifying the members of these groups is important. A list is accurate if the telephone numbers on the list actually do contain members of the target group. If the list is inaccurate, then a larger screener cost is incurred.

3. SAMPLING HOUSEHOLDS

This chapter describes the sample design and selection of households for CHIS 2003. We begin by defining the target population and the persons included and excluded in the survey. Target numbers of completed adult interviews by county and for the supplemental samples are then described. The remainder of the chapter describes how the sample of telephone numbers was selected in order to achieve the stated goals. The last section reviews the statistical issues considered in arriving at the allocation of the sample by county.

3.1 Population of Interest

The CHIS 2003 sample was intended to represent the adult (18 and older) residential population of California, as well as adolescents (aged 12-17) and children (aged 11 and under). Eligible residential households included houses, apartments, and mobile homes occupied by individuals, families, multiple families, extended families or multiple unrelated persons, provided that the number of unrelated persons was less than nine. Persons living temporarily away from home were eligible and enumerated at their usual residences. These include college students in dormitories, patients in hospitals, vacationers, business travelers, and so on. The survey excluded group quarters, – any unit occupied by nine or more unrelated persons (e.g., communes, convents, shelters, halfway houses, or dormitories). Institutionalized persons (e.g., those living in prisons, jails, juvenile detention facilities, psychiatric hospitals and residential treatment programs, and nursing homes for the disabled and aged), the homeless, persons in transient or temporary arrangements, and those in military barracks were also excluded. As described in Chapter 2, some individuals who were part of the residential population did not have a chance of selection, including those living in households without landline telephones (either without any telephone service or with cellular telephone service only), and children and adolescents living in a household without a parent or legal guardian.

3.2 Sample Allocation

In this section we describe the targeted number of completed interviews for CHIS 2003. We begin by discussing the RDD sample and then deal with the supplemental samples.

Two of the goals of CHIS 2003 were (1) to produce reliable statewide estimates for the total population in California and for its larger race/ethnic groups, as well as for several smaller ethnic groups (i.e., Koreans and Vietnamese), and (2) to produce reliable estimates at the county level for as many counties as possible. These goals required a compromise in allocating the sample. To achieve the most reliable statewide estimates, the optimal design is to allocate the sample to counties proportional to their population. On the other hand, the optimal allocation for producing county-level estimates is to assign each county an equal sample size. In this section we present the final compromise that allowed for both precise statewide estimates and reliable county-level estimates for most of the counties in California. We also discuss the rationale for the stratification and sample allocation, but we leave the more detailed statistical issues until a later section.

The 58 California counties were grouped into 41 strata as shown in Table 3-1. These strata are the same as in CHIS 2001. Thirty-three of the 35 counties with a population of 100,000 or more correspond to individual sampling strata. The two remaining counties with over 100,000 persons are each combined with an adjoining smaller county to form a stratum. The 23 remaining counties with populations of less than 100,000 were grouped geographically into six strata for analytic reasons.

Because of the need to produce reliable estimates for the counties, the sample allocation is not in all cases proportional to the population across counties. With a proportional allocation, the estimates from the moderate and smaller counties would be based on small sample sizes and would not be adequate for the envisioned analyses. To achieve the goal of producing local or county estimates, the sample sizes from the largest counties are re-distributed to the smaller counties. The target sample sizes ranged from 10,084 in Los Angeles to 400 in the smaller strata. The minimum target sample size of 400 completed adult interviews was set for each stratum. The RDD target goals are shown in Table 3-1. The goals in Table 3-1 include within-county supplemental samples in the Special Planning Area (SPA) of Antelope Valley in Los Angeles County, and in the cities of Hayward and Oakland in Alameda County.

CHIS 2003 had a goal for the RDD sample of completing 40,000 adult interviews, between 3,000 and 4,000 adolescent interviews (depending on compliance since parental consent and adolescent agreement are required), and from 8,000 to 10,000 child interviews conducted with knowledgeable parents or guardians. The RDD goal for adult interviews in CHIS 2003 was approximately 15,000 interviews lower than that for CHIS 2001.

Table 3-1. Targeted number of complete adult interviews for the RDD sample by county

	Sampling stratum	Targeted number of adult interviews	Population Size
1	Los Angeles [a]	10,084	Over 9,000,000
2	San Diego	2,279	1,200,000 or greater
3	Orange	2,142	
4	Santa Clara	1,296	
5	San Bernardino	1,211	
6	Riverside	1,160	
7	Alameda[b]	3,989	
8	Sacramento	1,039	800,000 to 1,200,000
9	Contra Costa	800	
10	Fresno	600	500,000 to 800,000
11	San Francisco	800	
12	Ventura	600	
13	San Mateo	600	
14	Kern	500	
15	San Joaquin	500	
16	Sonoma	500	100,000 to 500,000
17	Stanislaus	500	
18	Santa Barbara	500	
19	Solano	500	
20	Tulare	500	
21	Santa Cruz	500	
22	Marin	500	
23	San Luis Obispo	500	
24	Placer	500	
25	Merced	500	
26	Butte	500	
27	Shasta	500	
28	Yolo	500	
29	El Dorado	500	
30	Imperial	500	
31	Napa	500	
32	Kings	500	

Table 3-1. Targeted number of complete adult interviews for the RDD sample by county (continued)

	Sampling stratum	Targeted number of adult interviews	Population Size
33	Madera	500	Small and medium counties combined
34	Monterey (pop. >100,000) San Benito (pop. <100,000)	500	
35	Humboldt (pop. >100,000) Del Norte (pop. <100,000)	500	
36	Siskiyou Trinity Lassen Modoc	400	Less than 100,000 population per county
37	Mendocino Lake	400	
38	Tehama Colusa Glenn	400	
39	Sutter Yuba	400	
40	Nevada Sierra Plumas	400	
41	Tuolumne Mariposa Calaveras Mono Amador Alpine Inyo	400	
	Total of 41 Strata	40,000	

[a] The Los Angeles sample includes the 250 additional cases in Antelope Valley

[b] The Alameda sample includes 2,790 additional cases.

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

At the beginning of the study, different allocations of the sample consistent with the budget constraints were evaluated. The UCLA CHIS staff consulted with some counties and other analytic groups to define the relative importance of particular types of estimates. Westat statistical staff helped evaluate each alternative and examined the consequences of the sample allocations. The main statistical issues were communicated by computing effective sample size for the main groups for the alternative designs. The expected effective sample size computations are discussed in Section 3.4.

Table 3-2 shows the sampling goals for completed adult interviews for Koreans and Vietnamese in CHIS 2003. These are the only two race-ethnic groups with statewide oversample goals in CHIS 2003. These groups are a subset of the seven race-ethnic groups oversampled in CHIS 2001. The table includes the expected number of Koreans and Vietnamese from the RDD sample and the surname samples. The surname list sample targets were adjusted during data collection as the actual RDD yield became known.

Table 3-2. Targeted number of complete adult interviews for the Korean and Vietnamese supplemental list samples

Subgroup	Targeted number of adult interviews		
	RDD	Supplement	Total
Korean	451	49	500
Vietnamese	375	125	500
Total	826	174	1,000

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

3.3 Stratification

In this section we describe the steps used in selecting the sample of telephone numbers for CHIS 2003. These steps include stratifying the telephone numbers by sampling strata, selecting the sample of numbers after adjusting for expected losses due to nonresponse, and subsampling the numbers based on mailable status and refusal status to improve the efficiency of the sample.

Since CHIS 2003 used a stratified sample, the first step was stratifying the sampling frame of 100-banks with one or more listed telephone numbers into non-overlapping strata corresponding to a ZIP code, a city, a county, or a group of counties. The geographic information required for stratification is available only at the exchange level[11], so 100-banks could not be assigned directly to a single stratum. All banks within an exchange were stratified indirectly by mapping the exchanges to a county represented by the stratum. However, some telephone exchanges actually service households in more than one county.

To solve the stratification problem, Genesys produced coverage reports for each county in California. The coverage reports listed all the exchanges in the county. For each exchange, the report gave the total number of listed households in the exchange and the proportion of listed households that are within the county. After combining the information of the coverage reports for all 58 counties, we created a frame of exchanges with variables for the number of listed households in each county that the exchange covers. Each exchange was assigned to the county that contains the most listed households. In CHIS 2003, there was also interest in obtaining a better sample distribution for Los Angeles County by Special Planning Areas (SPAs). Using ZIP code information, telephone exchanges in Los Angeles were classified into eight subsampling strata, each representing a SPA. Telephone exchanges that crossed SPAs were assigned to the SPA with the most listed households. There were no targets for individual SPAs, so the sample for Los Angeles was allocated proportionally by these substrata, except for the sample for

[11] A telephone exchange consists of 10,000 consecutive telephone numbers with the same first six digits including area code. An exchange is a set of area codes and prefixes serving the same geographic area.

Antelope Valley (SPA=1). The sample for Antelope Valley included an additional sample to yield 250 adult interviews more than what would be expected from the proportional allocation.

One month after the beginning of the data collection, the target sample size for Alameda County was initially increased by 801 interviews. The sample was proportionally allocated to Hayward, Oakland, and the remainder of Alameda. One month later the target samples for the cities of Hayward and Oakland were further increased by 990 and 940 interviews each, increasing their total sample sizes to 1,162 and 1,516 respectively. The new sample targets also included 200 African-American adult interviews in Hayward beyond the 141 expected cases from the 1,162 in Hayward. The final goal for Alameda was 3,989 adult interviews after the increases in sample.

The telephone exchanges were classified in substrata for areas defined by Hayward, Oakland and remainder of Alameda before the sample selection of the additional cases. However, telephone exchanges overlapping the substrata within Alameda County proved problematic. If exchanges were assigned to the substrata that had the most telephone numbers, then they would cover a large proportion of households outside the cities. If the substrata were created using exchanges that were contained within the cities, then the substrata covered only a proportion of the households in the cities. The creation of the substrata also had to consider the oversampling of African Americans in Hayward and the fact that telephone numbers had already been selected and fielded for Alameda County. After analyzing the distribution of the initial set of completed cases in Alameda County, we created nine substrata in Alameda County based on the concentration of households in the cities and African Americans in the telephone exchanges. Table 3-3 shows the definition of the substrata in Alameda county. The sample in these substrata was released sequentially depending on the number of completed interviews that had been achieved during data collection.

Table 3-3. Definition of substratum for the Alameda supplemental samples

		Concentration in Area				
Stratum	Substratum	Hayward	Oakland	Remainder of Alameda	Concentration African American Households	Designation
7	10711	High			High	HY-H-AA-H
7	10712	High			Low	HY-H-AA-L
7	10721	Low	Low		High	HY-OK-AA-H
7	10722	Low	Low		Low	HY-OK-AA-L
7	10731	Low		Low	High	HY-RA-AA-H
7	10732	Low		Low	Low	HY-RA-AA-L
7	20720		High			OK-H
7	30720		Low	Low		OK-RA-L
7	30730			High		RA-H

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

As mentioned in Chapter 2, we used disproportionate stratified sampling to oversample Koreans and Vietnamese without increasing the sample size allocated to any stratum (the stratum sample size was fixed). Although the geographic oversampling increased the RDD sample yield, the additional sample sizes for Koreans and Vietnamese were not large enough to achieve the CHIS 2003 goals for these groups. In order to achieve the desired sample goals, the sample design also contemplated the use of surname lists to supplement the RDD sample.

Once the primary sampling strata were created, we examined the concentration of Koreans and Vietnamese in the areas covered by the telephone exchanges within strata. Using ZIP code level information from Census 2000 we classified the telephone exchanges into high/low concentration substrata using different cut-off points (i.e., *x* percentage or more Koreans or Vietnamese residing in the telephone exchange). The telephone numbers in the high-density substrata would be sampled at a rate that was *y* times that in the low-density substrata. Because the sampling rates for the two substrata were constrained, the total yield for the stratum was not affected. As part of our analysis, we examined different sampling rates and cut-off points for the creation of the substrata. We computed expected nominal sample sizes, design effects, and effective sample sizes for these designs. The examined cut-off points for the creation of substrata varied from four to eight percent (of Koreans or Vietnamese) and the examined range for the ratio of sampling rate for the high density to the low density strata varied from one (no oversampling) to five. We also examined the effect on the nominal sample size and effective sample size for other race-ethnic groups such as Chinese, Japanese, Filipino, Asian, African Americans, Latinos, and American Indians. Figures 3-1 and 3-2 show the percentage increase in expected and effective sample size for Koreans and Vietnamese at different oversampling rates and cut-off points for the creation of the high and low concentration substrata (four percent or more, five percent or more, and six percent or more). Figure 3-3 shows the design effect for some of the other Asian groups as a function of oversampling rates for substrata created using 6 percent or more. Figure 3-4 shows the effect oversampling Korean and Vietnamese in these substrata had on the expected and effective sample sizes for Chinese, Japanese and Filipinos.

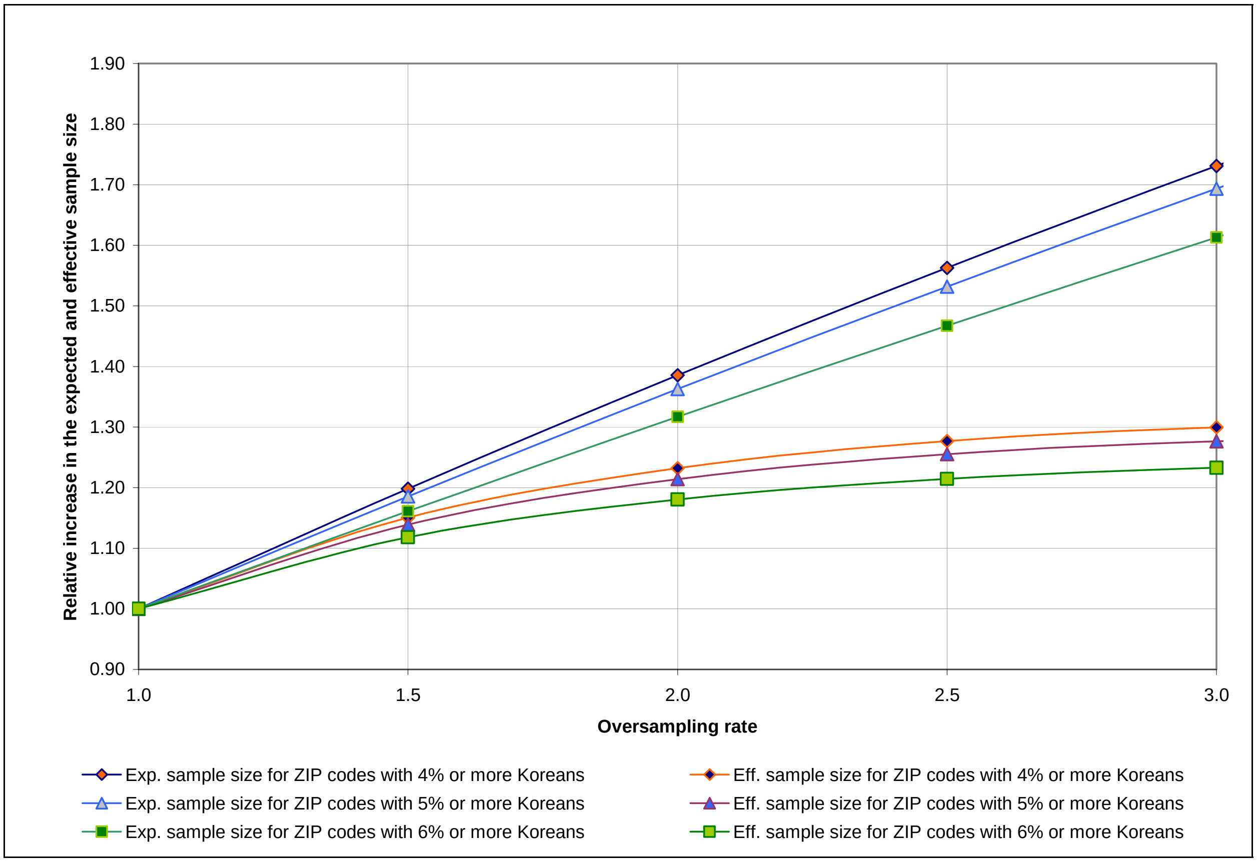


Figure 3-1. Relative increase in expected and effective sample sizes for Koreans when oversampling ZIP codes with a higher percentage of Korean and Vietnamese.

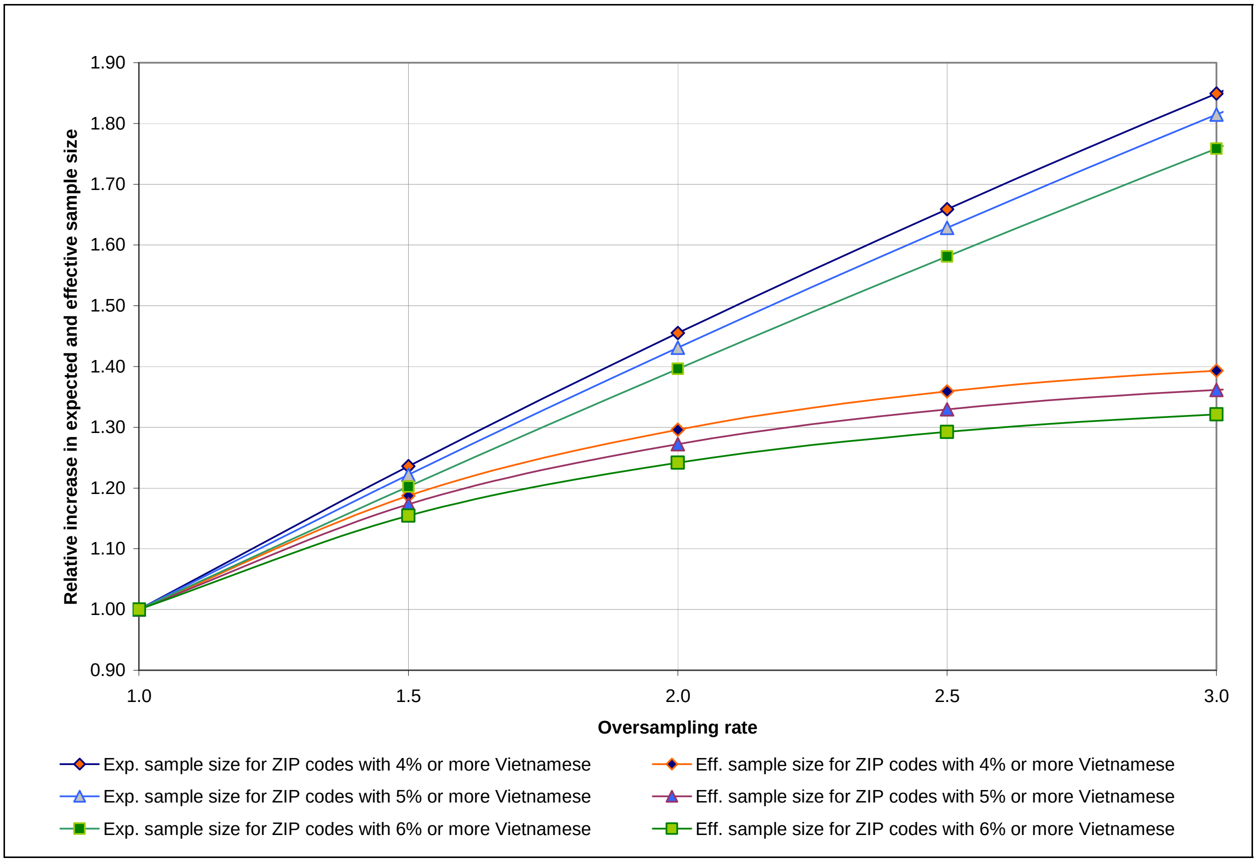


Figure 3-2. Relative increase in expected and effective sample sizes for Vietnamese when oversampling ZIP codes with high percentage of Korean and Vietnamese.

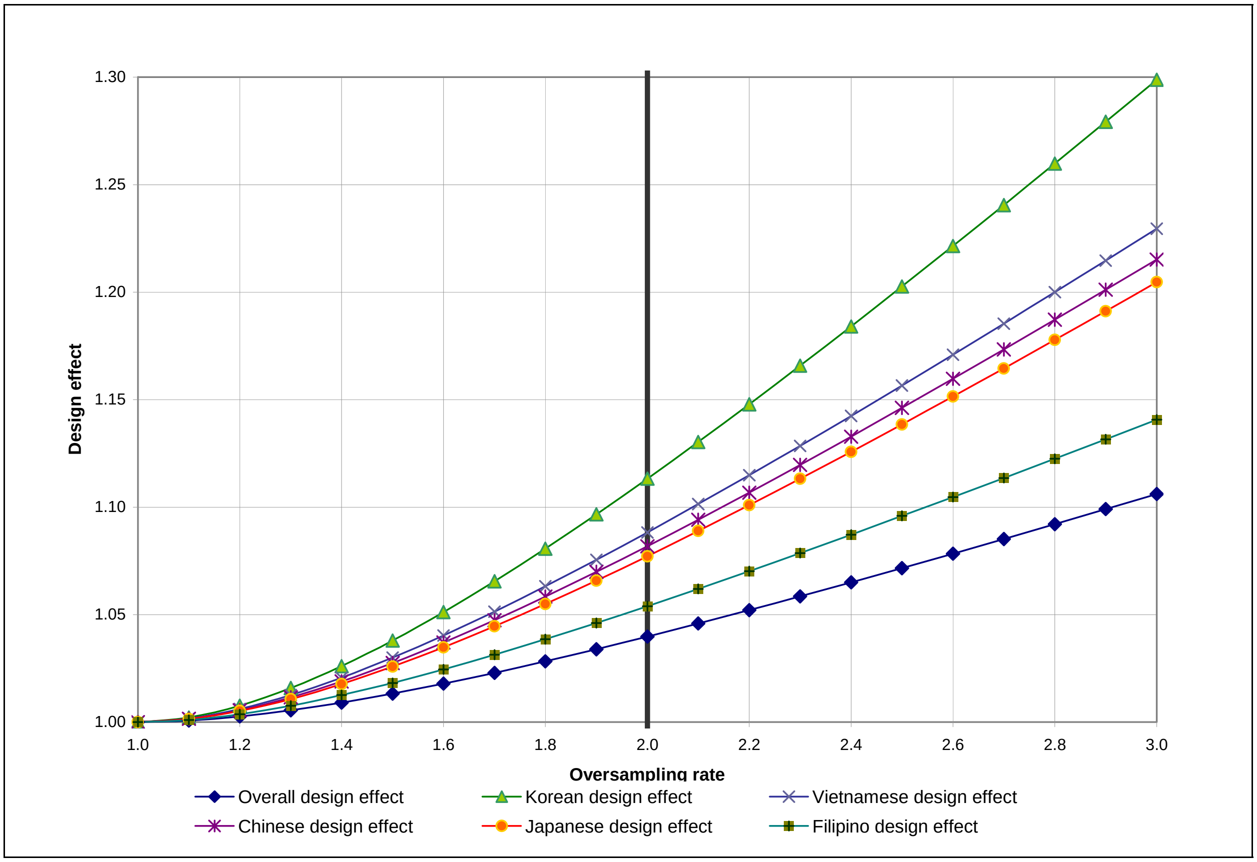


Figure 3-3. Design effects for different Asian ethnic groups when oversampling ZIP codes with a high percentage of Korean and Vietnamese (6 percent or more).

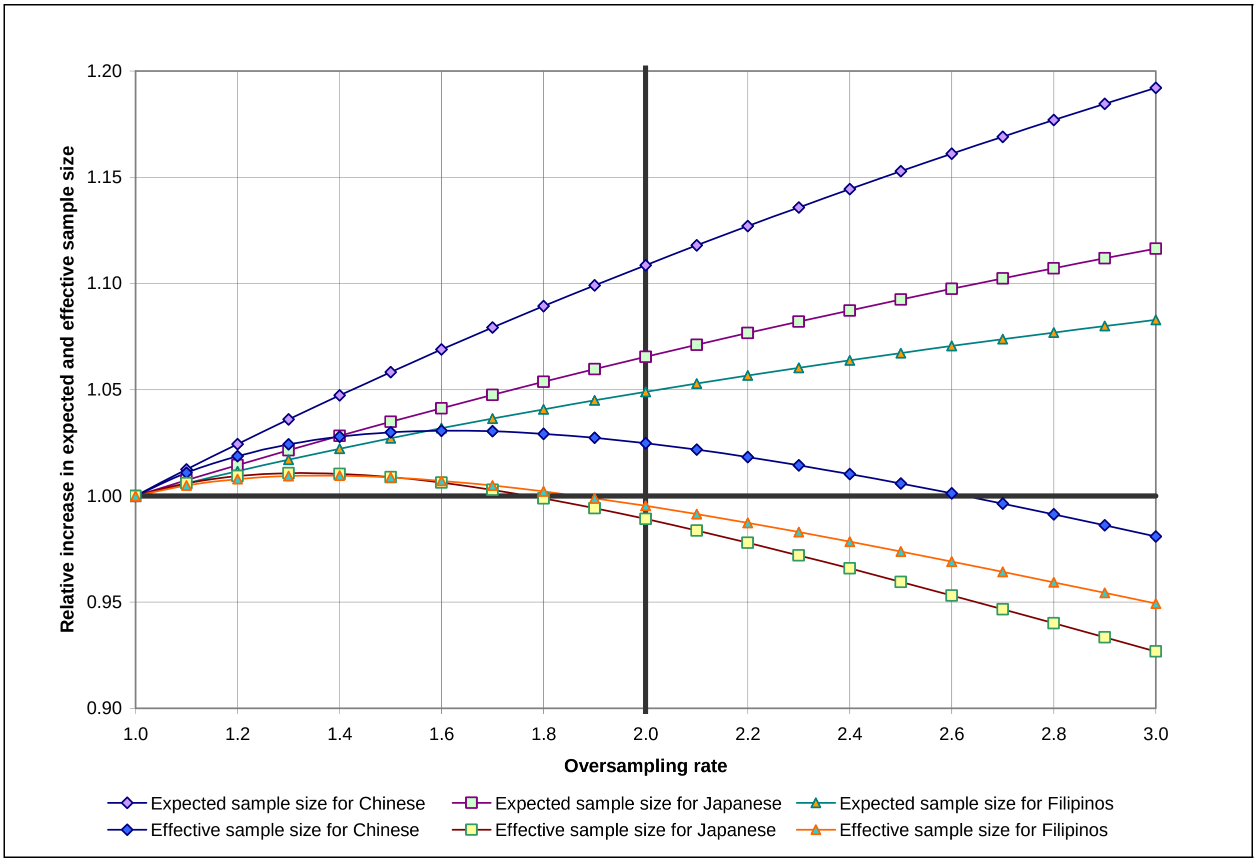


Figure 3-4. Relative increase in expected and effective sample sizes for Asian groups when oversampling ZIP codes with a high percentage of Korean and Vietnamese (6 percent or more).

We determined that a cut-off point for defining the high-concentration areas of six percent or more Koreans and/or Vietnamese combined and sampling the high concentration areas at twice the rate of the low concentration areas were near the optimal levels. This cut-off and sampling rate increased the expected yield of Koreans and Vietnamese and did not inordinately inflate the design effect nor decrease the effective sample sizes for the Asian groups of interest that are not oversampled.

The creation of high and low concentration substrata was restricted to four sampling strata where the Korean and/or Vietnamese population was large enough to produce increases in the expected number of interviews. The sampling strata covered approximately 78 percent of the Korean and Vietnamese population in California while the oversampled exchanges represented less than 40 percent of the Korean and Vietnamese population. Table 3-4 shows the four sampling strata, the high and low-concentration substrata, the number of households and the percentage of the households within each substratum.

Table 3-4. Stratum definitions and sample sizes for strata with high and low Korean and Vietnamese substrata

Stratum	County	Korean and Vietnamese population density substratum*	Number of households	Percentage
1	Los Angeles	High	457,370	15
		Low	2,676,404	85
		Total	3,133,774	100
2	San Diego	High	90,200	9
		Low	904,477	91
		Total	994,677	100
3	Orange	High	335,477	36
		Low	599,810	64
		Total	935,287	100
4	Santa Clara	High	233,886	41
		Low	331,977	59
		Total	565,863	100

* High density areas were defined as those with 6 percent of more Koreans or Vietnamese in the exchange

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

Table 3-5 shows the effect of this disproportionate sampling on race-ethnic groups other than Koreans and Vietnamese. The table shows the percentage increase in sample size and net increase in design effect resulting from disproportionate sampling compared to a design with no disproportionate sampling. Large gains in sample size for Koreans and Vietnamese were expected, without incurring large increases in the design effect. There were some losses in expected sample size for some ethnic groups, in particular African-Americans. Still, the African-American sample was large enough for precise estimates at the state level and county levels. Despite the increased design effect for Koreans and Vietnamese, this design effect was still smaller than that for a design where the populations were oversampled using only surname lists.

Table 3-5. Increase in expected sample size and design effect compared to a design without disproportionate sampling by race-ethnic groups*

Race/Ethnicity Group	Increase in expected sample size (Percentage)	Increase in design effect
Korean[a]	27.65	0.13
Vietnamese [a]	21.67	0.06
Chinese	10.73	0.08
Japanese	6.71	0.10
Filipino	4.16	0.06
African American	-5.54	0.02
Hispanic	-4.43	0.04
American Indian	-2.33	0.04

[*] Negative numbers imply increases in effective sample size

[a] Based on RDD sample only

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

Another concern was the effect of this disproportionate sampling on the sample size in Los Angeles County. As mentioned earlier, a goal was to produce estimates at the SPA level in CHIS 2003. Although the SPAs are not considered as separate sampling strata, disproportionate sampling to increase the yield of Koreans and Vietnamese could have drastically reduced the expected sample size for the larger SPAs. However, analysis showed that disproportionate sampling had relatively little effect on the effective samples sizes for the SPAs, with the largest decrease in Antelope Valley and South as shown in Table 3-6.

Table 3-6. Decrease in expected effective sample size in SPAs compared to a design without disproportionate sampling

Special Planning Area (SPA)	Decrease in effective sample size * (Percentage)
1. Antelope Valley[a]	12.7
2. San Fernando	2.8
3. San Gabriel	-5.5
4. Metro	-17.4
5. West	7.8
6. South[b]	12.7
7. East	2.1
8. South Bay	0.8

[*] Negative numbers imply increases in effective sample size

[a] Excludes supplemental sample of 250 adults

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

The other procedure for oversampling Koreans and Vietnamese used surname lists. The sampling frames for these supplemental samples were created by MSG using surnames likely to be of Korean and Vietnamese origin for the state of California. By matching the surnames for the subgroup against the listed surname in the White Pages for the state, a sample was selected for each subgroup. For both subgroups, the sampling was done over the entire state. Table 3-7 shows the size of the surname lists. It also shows that there was a sizable overlap between the two lists.

Table 3-7. Number of records in the surname frames

Surname frame	Number of records
Korean only	129,127
Vietnamese only	96,710
Korean and Vietnamese	81,604

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

3.4 Sample Selection

The number of telephone numbers selected in any RDD survey has to be greater than the targeted number of completed interviews to account for a variety of factors. For example, a substantial percent of the sampled telephone numbers is not residential. For CHIS 2003 the sample of telephone numbers was inflated to deal with the losses due to following sources:

- Nonworking, nonresidential, and never answered numbers;
- Subsampling by mailable status;
- Subsampling for refusal conversion
- Nonresponse to screening interview; and
- Nonresponse to extended adult interview.

The first, fourth, and fifth sources noted above are typical of all RDD surveys. To deal with these losses we used information from CHIS 2001 to estimate the percentage of the telephone numbers that would not be residential and the percentage that would not respond to the screener and extended interviews, and increased the sample size accordingly. The only sources of loss that require additional discussion are the subsampling by mailable status and for refusal conversion. As mentioned in Chapter 2, to increase the efficiency of the CHIS 2003 RDD sample we first stratified the telephone numbers within sampling stratum by mailable status and then subsampled those numbers that were not mailable. The subsampling rate was 75 percent, meaning that about 25 percent of the telephone numbers without a mailable address were removed from the sample. During CHIS 2003 sample selection, 60 percent of the telephone numbers were flagged for refusal conversion. Refusal conversion efforts were made only to flagged telephone numbers after the respondent refused to do the screener interview. Taking all of these factors into consideration, a sample of 459,200 telephone numbers[12] was selected for CHIS 2003. The data collection procedures are discussed in Report 2: Data Collection Methods.

The last samples selected were the surname list samples. Calculating the number of telephone numbers needed to meet the goals for the Korean and Vietnamese subgroups required estimating the losses from nonresponse and nonhousehold numbers as with the RDD sample, as well as from eligiblity (households without a member of the target ethnic group). The expected loses from these sources were estimated using results from the list samples in CHIS 2001, and the sample size was increased accordingly. Table 3-8 shows the total number of telephone numbers drawn for the supplemental list samples.

Table 3-8. Number of telephone numbers drawn for the supplemental list samples

Surname Frame	Sample size
Korean	2,158
Vietnamese	1,667
Total	3,825

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

Table 3-9 summarizes the size of each type of sample. The total sample size for CHIS 2003 including both the RDD and supplemental samples was 463,025 telephone numbers.

[12] It excludes telephone numbers drawn form the supplemental list samples

Table 3-9. Number of telephone numbers sampled by type of sample[*]

	Sample Size
RDD Sample	345,700
Antelope Valley	200
Hayward	22,918
Hayward African Americans[a]	12,387
Oakland	20,405
Remainder of Alameda	57,590
Korean surname list[b]	2,158
Vietnamese surname list[b]	1,667
Total	463,025

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

[*] Duplicate telephone numbers were removed. RDD numbers sampled in other subsamples were counted as RDD records.

[a] Sample in selected telephone exchanges that covered ZIP codes with high concentration of African Americans.

[b] Not sampled by separate strata.

3.5 Expected Design Effect

Sections 3.2 and 3.3 described the allocation of the sample of telephone numbers by sampling stratum and substratum and noted that it was a compromise among three goals: to produce reliable estimates for the entire state, to produce estimates at the county level, and to oversample Koreans and Vietnamese. Allocating the sample proportionally to the population in the counties would be approximately optimal for statewide estimates. For county estimates, an equal allocation would be more efficient. In this section, we describe the statistical methods used to examine the efficiency of the sample under different allocations. These methods were used to help guide the sample allocation for CHIS 2003.

If CHIS 2003 had been a simple random sample, it would be relatively simple to predict the precision of the estimates. Under the assumption of simple random sampling, suppose we wish to estimate a proportion of adults with a characteristic, say *p*. If the sample size is large enough, then the standard (1-α)·100% confidence interval of the estimated proportion is

$$\left(p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}, p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \right) \qquad (1)$$

where $z_{1-\alpha/2}$ is the critical value from the standard normal distribution, and *n* is the number of completed interviews. This form of the confidence interval is not appropriate for CHIS 2003 for several reasons. The main reason we discuss below is because the allocation of the sample to the counties does not produce a simple random sample. The other reasons that (1) is not fully appropriate are sampling within households and other adjustments to the estimation weights. These issues are covered in Report 5: Weighting and Variance Estimation.

To adjust (1) to account for the sample allocation to the counties or strata we introduce the concept of a design effect. Kish (1992) discusses the design effect in some detail. Here we simply note that in stratified designs like CHIS, the design effect measures the departures with respect to a sample proportionally allocated among the strata. A sample with proportional allocation has a design effect of one. Departures from proportional allocation result in design effects greater than one.

The design effect due to departures from proportional allocation can be computed as

$$D = \left(\sum_{h=1}^{H} W_h k_h \right) \left(\sum_{h=1}^{H} \frac{W_h}{k_h} \right), \tag{2}$$

where W_h is the proportion of the population in sampling stratum h computed as $W_h = N_h \left(\sum N_h \right)^{-1}$, where N_h is the population total in stratum *h*, and k_h is the relative sampling rate for strata *h*. More specifically, k_h is defined as $k_h = \frac{n_h}{N_h} \frac{N_1}{n_1}$, where n_h is the sample size in stratum *h* and the reference stratum is set to be stratum 1 so that $k_1 \equiv 1$(the choice of the reference stratum does not affect the computations since the relative sampling rates are the only factors involved).

Using the design effect computed in this way, we can estimate the effective sample size for a stratified sample with a given allocation. The effective sample size is the number of cases needed from the stratified sample to produce estimates with the same precision that would be expected from a simple random sample design. The effective sample size n_{eff} is computed as

$$n_{eff} = \frac{n}{D}. \tag{3}$$

where n is the nominal sample size and D was defined above.

In CHIS 2003, we expected to complete 39,941[13] adult interviews from the RDD sample. The Hayward African American supplemental sample and the Korean and Vietnamese supplemental list samples were not included in this evaluation. The expected nominal sample sizes (the number of adult

[13] Excludes the 59 additional African American cases in Hayward

interviews), the expected design effects due to the sample allocation to the strata using (2), and the expected effective sample sizes using (3) are given in Table 3-10. The expected design effects and effective sample sizes are given for the entire state and for domains defined by race and ethnicity. It is important to remember that the design effects are computed at the household level and they do not include any adjustments for nonresponse, within-household sampling, or other weighting adjustments.

Table 3-10. Expected design effects and effective adult sample size associated with the sample allocation

	Domain	Expected nominal sample size	Expected design effect	Expected effective sample size
1	White	26,809	1.25	21,426
2	Native Hawaiian/Pacific Islander	105	1.32	79
3	African-American[a]	2,879	1.27	2,275
4	American Indian/Alaskan Native	416	1.39	299
5	Asian	3,843	1.22	3,161
6	Other (One Race)	4,353	1.22	3,559
7	Two or More Races	1,462	1.26	1,163
8	Overall	39,941	1.25	31,909

[a] Excluding the supplemental sample of Hayward African Americans

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

For example, the expected yield from the CHIS 2003 sample for African-Americans was 2,879 adults[14]. Due to the allocation of the sample, the expected effective sample size was 2,275. The 95 percent confidence interval for an estimated proportion can be computed by using the entries in this table and replacing n in (1) by n_{eff} . For example, for estimating a proportion of p = 0.5 for American Indian/Alaska Natives, the 95 percent confidence interval is

$$\left(0.5-1.96\sqrt{\frac{0.5^2}{299}},0.5-1.96\sqrt{\frac{0.5^2}{299}},\right)=\left(0.4433,0.5567\right)$$

As the UCLA CHIS staff consulted with various groups in California to evaluate the data needs that CHIS could help to support, they developed different allocation schemes for distributing the sample to the counties. The effects of these allocations were examined by using the methods presented above. The UCLA CHIS staff then chose the sample allocation that best satisfied the needs of survey data users.

[14] It excludes the supplemental sample of Hayward African Americans

4. WITHIN-HOUSEHOLD SAMPLING

Once the sample of telephone numbers is selected, interviewers call the numbers, select and conduct interviews with sampled persons within the household. This chapter describes the procedures for selecting the sample of persons within households for CHIS 2003. Person subsampling was done primarily to reduce respondent burden at the household level. Samples of adults, children, and adolescents within the household were selected using different sampling procedures, but one adult, and up to one child and one adolescent were sampled within each household. The within-household sampling procedures were developed to maximize the analytic utility of the data collected from the respondents. The next section describes the within-household sampling alternatives we evaluated to accomplish this and the reasons for choosing the specific method of sampling. The second section describes sampling adults within sampled households. The third section gives the methodology used for sampling children and adolescents. The last section details how the within-household sampling was implemented in CHIS 2003.

4.1 Sampling Alternatives

The general idea for the CHIS 2003 sample design was to sample randomly one adult from all the adults in each sampled household. In addition, in those households with adolescents (ages 12-17) and/or children (under age 12), one adolescent and one child were to be sampled and interviewed (a parent of the child was interviewed about the child). One approach to accomplishing this goal is simply to list all the persons in the age group (adult, child, and adolescent) in the household and select one person randomly from each group. We call this the *completely random* sampling method.

The completely random sampling method is not a problem in most households because most households have only one family. However, in households with two or more families, the completely random method could result in selecting persons from the different age groups who were not members of the same family. This situation is undesirable because the adult interview collected data about the family of the sampled adult. The data from the adult interview are of great value for the analysis of the data from the child and adolescent interviews. If the sampled child and/or sampled adolescent were not members of the same family as the sampled adult, then the data collected about them would be of very limited utility.

To illustrate this type of household consider Figure 4-1. It shows the familial relationships in a household with two families (*F1* and *F2*). In the figure, family *F1* consists of 3 adults, (*AD1*, *AD2* and

AD3) and one adolescent (*TN1*); *AD3* is a young adult (18 or older) child of *AD1* and *AD2*. A second family, *F2*, shares the same household but the members of *F2* are not related to the family *F1*. Family *F2* consists of one adult *AD4* and one adolescent *TN2*.

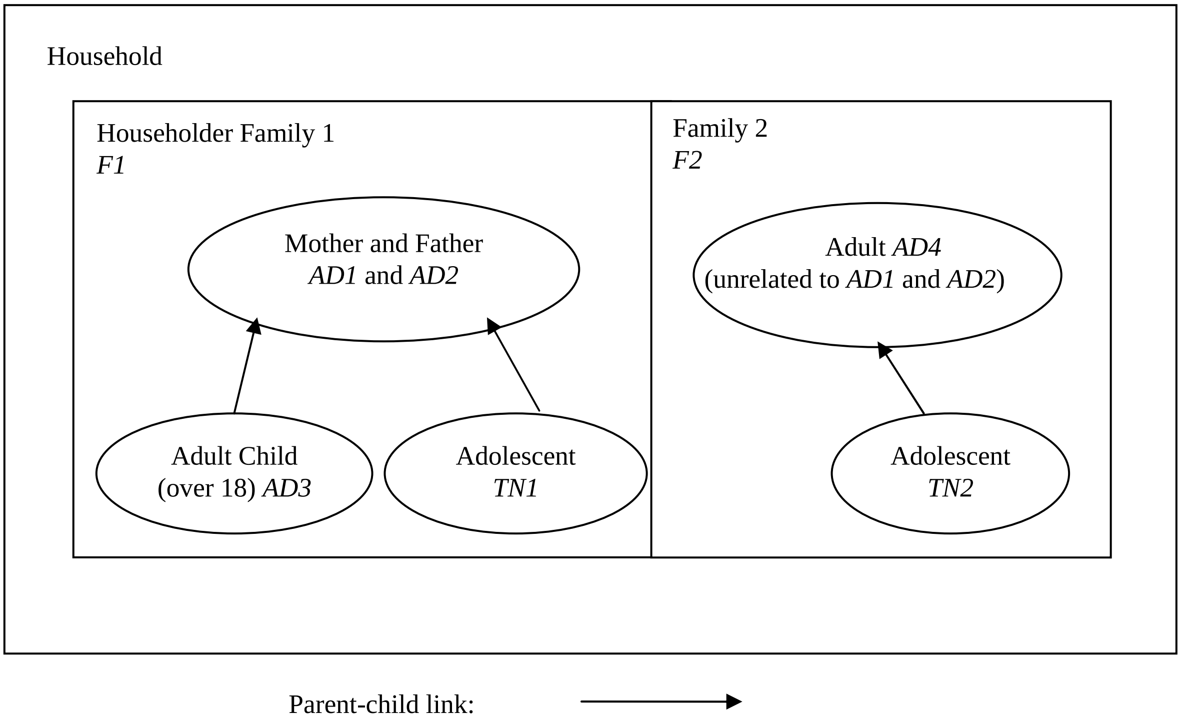


Figure 4-1. Illustrative household with two families

If one adult and one adolescent were selected using the completely random method, one possible outcome is the selection of adult *AD4* and adolescent *TN1*. In this case, the family data collected from the *AD4* would not be useful for describing the family circumstances of *TN2* because they are not members of the same family.

To resolve this analytic problem, a second sampling alternative was adopted for CHIS 2003. We call this method the *linked* sampling approach. In this approach, the children and adolescents in the household were linked to the adults. Children and/or adolescents for whom a sampled adult was a blood, adoptive, or foster parent or other legal guardian were considered as linked or "associated" with that adult.

In the linked sampling method persons are sampled in two phases. In the first phase, an adult is randomly sampled from all the adults in the household. In the second phase, a child and/or adolescent is sampled from all the children/adolescents associated with the sampled adult. In the example in Figure 4-1, if adult *AD4* is sampled, then the only adolescent eligible for sampling is *TN2* and that adolescent would be selected. Since the sampling of adolescents (and children) is a two-phase procedure, the probability of sampling the adolescent is the product of the probability of sampling the adult (phase one) and the probability of sampling the adolescent from the all the adolescents associated with that adult (phase two).

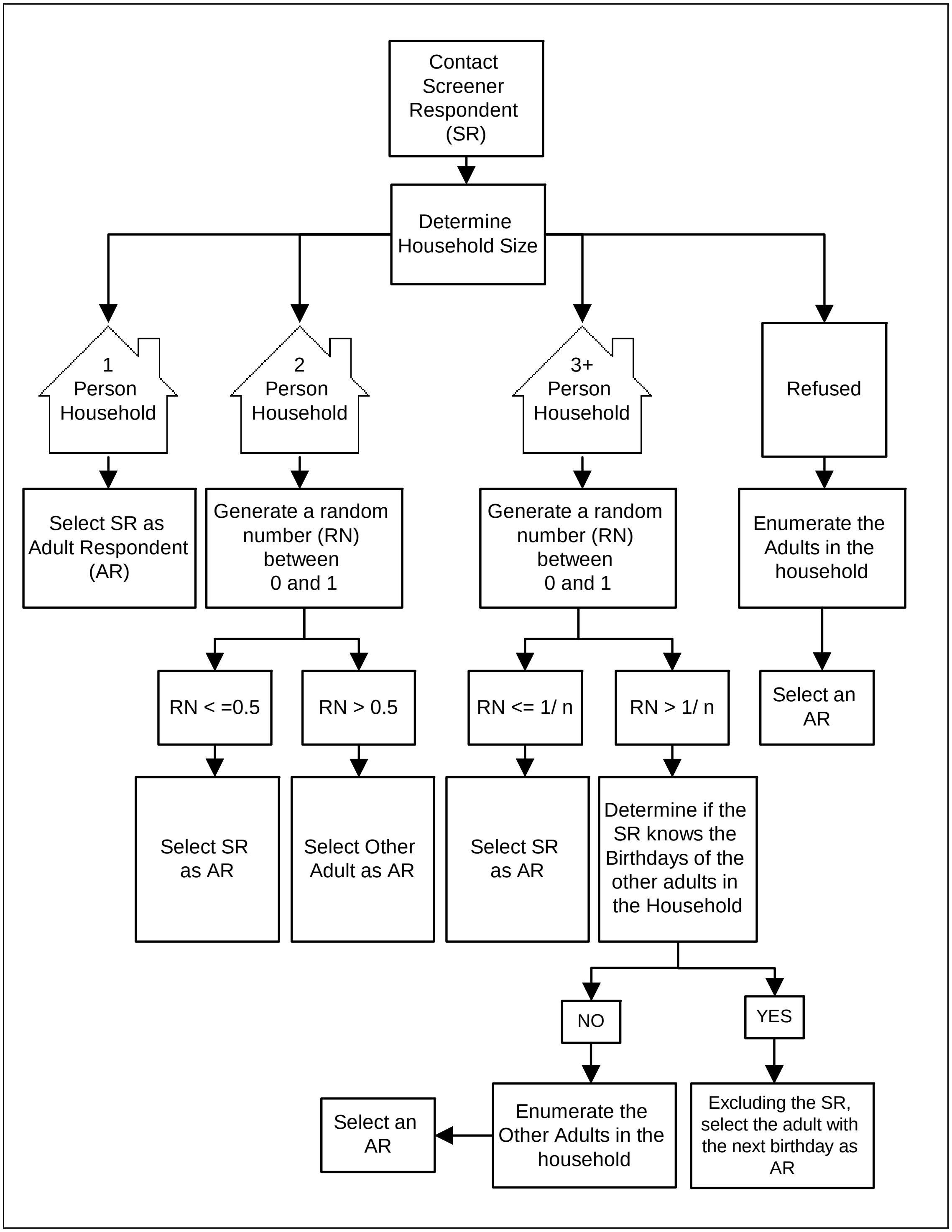
To use the linked sampling method, data are needed linking children and adolescents in a household to the sampled adult and his/her spouse (children or adolescents linked to both the sampled adult and spouse could be selected if either adult was sampled). These data were collected in the adult interview in CHIS 2003. We expected that in a very few households it would not be possible to link or associate a child or adolescent to an adult because of unusual household structures. A child or adolescent not associated with an adult would not have a chance of being selected. In CHIS 2003, the UCLA Institutional Review Board (IRB) directed that only children and adolescents of the sampled adult could be selected. Therefore, unassociated children and adolescents in a household could not be randomly linked to an adult in the household as in CHIS 2001. Based on the results from CHIS 2001, only 17 of 16,523 (0.10 percent) households with children had any unassociated children; of the 10,867 households with at least one adolescent only 37 (0.34 percent) had at least one unassociated adolescent. The bias due to excluding unassociated children and adolescents in CHIS 2003 resulting from this restriction was expected to be very small. Due to changes in the way adults, children and adolescent were enumerated in CHIS 2003, we were unable to determine the number of unassociated children and adolescents in the sampled households.

4.2 Sampling Adults

In CHIS 2003, an adult is defined as any person 18 years or older residing in the household. The procedure to select adults in CHIS 2003 was different from the one used in 2001. In CHIS 2001, one adult per household was sampled using the Kish method with full enumeration of adults in the household (Kish, 1949). Although in most cases adults were sampled with equal probability in CHIS 2001, some adults were selected with differential probabilities under special conditions. In households with both adults younger than 24 years old and adults 40 years old or older, adults 40 years old or older had twice the chance of being selected. This method was used in order to reduce the chances of selecting adult children, thereby including more children and adolescents in the survey.

In CHIS 2003, a new approach called the Rizzo method (see Rizzo et. al., 2004 for a complete discussion of the advantages of the method and its implementation) was used to sample adults in the household. The advantage of this method is that the enumeration of adult household members is bypassed in most households, so it is less intrusive but results in a valid probability sample. In this method, all sampled adults have an equal probability of selection. A sampled adult is selected using the following algorithm (see Figure 4-2). The steps are as follows:

- Ask the screener respondent (who must be an adult living in the household) how many eligible adults are in the household. The respondent answers N=1, 2, 3, ...;
- If there is only one eligible adult in the household, then that adult is selected;
- If there are two eligible adults in the household, then the CATI system accesses a pre-generated uniform random number between 0 and 1.
    - If the random number is less than or equal to 0.5 then the screener respondent is selected;
    - If the random number greater than 0.5 then the other adult is selected;
- If there are more than two eligible adults in the household, then the CATI system accesses a pre-generated uniform random number between 0 and 1.
    - If the random number is less than or equal to the inverse of the number of eligible adults in the household then the screener respondent is selected;
    - If the random number is greater than the inverse of the number of eligible adults in the household then, then the screener respondent is asked which of the eligible adults is the next to have a birthday; and
        - If the screener respondent knows which eligible adult is next to have a birthday, then the adult with the next birthday is selected.
        - If the screener respondent does not know which eligible adult is next to have a birthday then the respondent is asked to list the eligible adults in the household (excluding the screener respondent) and the CATI system randomly chooses one of the adults from this roster.
- If the number of eligible adults in the household is unknown then the screener respondent is asked to list the eligible adults in the household (including the screener respondent) and the CATI system randomly chooses one of the eligible adults from this roster. No other sampling steps are necessary.



Contact Screener Respondent (SR)
Determine Household Size
1 Person Household
2 Person Household
3+ Person Household
Refused
Select SR as Adult Respondent (AR)
Generate a random number (RN) between 0 and 1
Generate a random number (RN) between 0 and 1
Enumerate the Adults in the household
RN < =0.5
RN > 0.5
RN <= 1/ n
RN > 1/ n
Select an AR
Select SR as AR
Select Other Adult as AR
Select SR as AR
Determine if the SR knows the Birthdays of the other adults in the Household
NO
YES
Select an AR
Enumerate the Other Adults in the household
Excluding the SR, select the adult with the next birthday as AR

Figure 4-2. Sampling an Adult in CHIS 2003

4.3 Child and Adolescent Sampling

The sampling for children and adolescents took place after the adult was sampled and completed the enumeration of all persons under 18 years old in Section G of the adult extended interview. If there were any children under 12 in the household who were associated with the sampled adult, then exactly one child was sampled and each associated child had an equal probability of selection. The same procedure was followed for sampling exactly one adolescent with equal probability from all the adolescents associated with the sampled adult.

As described in Section 4.1, children or adolescents not associated with the sampled adult in a household were not eligible to be selected in this second phase of sampling. In some such cases, the sampled adult did not have any associated child or adolescent. Consequently, some households with a child or adolescent had none sampled.

In CHIS 2001, adults were sampled with different probabilities of selection in order to reduce the chances of selecting adult children, thereby increasing the chance of including children and adolescents in the survey. In contrast, in CHIS 2003 adults were sampled with equal probability because of the use of the Rizzo method. To evaluate the effect of the changes of sampling procedures for adults and its impact on the number of children and adolescent in CHIS 2003, we tabulated the results of the sampling procedures for children in Table 4-1 and for adolescents in Table 4-2. Table 4-1 shows the number of households with children where a child was selected at different stages of the extended interview. Children were sampled in Section H of the adult extended interview in CHIS 2001 and Section G in CHIS 2003. The table shows the number and percentages of households where an adult and child extended interviews were completed.

Table 4-1. Number and percentage of selected/Not Selected Children in households with children in CHIS 2001 and CHIS 2003

Households with children	Interview completed through child selection section*				Completed Adult Interview				Completed Child interview	
	2001		2003		2001		2003		2001	2003
Child selected	14,338	91.8%	10,440	87.9%	14,254	91.8%	10,106	87.9%	12,593	8,526
Child not selected	1,279	8.2%	1,444	12.2%	1,271	8.2%	1,400	12.2%	0	0
Total	15,617	100.0%	11,884	100.0%	15,525	100.0%	11,506	100.0%	12,593	8,526

* Adult Extended interview section H in CHIS 2001 or section G in CHIS 2003.

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

Table 4-2. Number and percentage of selected and not selected adolescents in households with adolescents in CHIS 2001 and CHIS 2003

Households with adolescents	Interviews completed through adolescent selection section*				Completed Adult Interviews				Completed Adolescent interviews	
	2001		2003		2001		2003		2001	2003
Adolescent selected	9,245	89.5%	6,857	82.1%	9,195	89.5%	6,655	82.1%	5,801	4,010
Adolescent not selected	1,088	10.5%	1,498	17.9%	1,085	10.6%	1,452	17.9%	0	0
Total	10,333	100.0%	8,355	100.%	10,280	100.%	8,107	100.%	5,801	4,010

* Adult Extended interview section H in CHIS 2001 or section G in CHIS 2003.

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

In CHIS 2003, there were 11,884 households with children where the sampled adult completed the extended interview through Section G. Among these, there were 1,444 households with children (12.2 percent) where a child was not sampled because the sampled adult was not the parent or legal guardian of any children in the household. This proportion is 3.9 percentage points lower in CHIS 2003 than in CHIS 2001. Similarly, Table 4-2 shows 8,355 households with adolescents where the sampled adult completed through Section G. Among these, there were 1,452 (17.9 percent) households where an adolescent was not sampled. This proportion is 7.3 percentage points higher in CHIS 2003 than in CHIS 2001. Thus the change of method for sampling an adult in CHIS 2003 had a slight effect on the number of number of sampled children, and a greater impact in the number of sampled adolescents. Of course, one advantage of this method is that the design effect due to differential sampling of adults that was present in CHIS 2001 was eliminated because all adults in a household had an equal probability of selection. The decrease in the design effect in the 2003 procedures also applies to the samples of children and adolescents.

4.4 Enumeration, Assignment, and Sampling Procedures

As described in the previous sections, the sampling of persons in CHIS 2003 was conducted in two phases, with adult sampling in the first phase and child and adolescent sampling in the second phase. The adult was sampled during the screener interview and the child and adolescent were sampled after the persons under 18 years old were enumerated at the end of Section G in the adult extended interview. We begin by giving the specific sampling procedures used and then conclude the section by discussing the overall probability of selection for each sampled person.

The following steps are the details of the selection process used in CHIS 2003.

1. Select an Adult Respondent (AR) using the Rizzo method as described in Section 4.2.
2. As part of the adult extended interview with the AR, the adult identifies if they have a spouse or partner (ARSP) living in the household. They also enumerate all children (ages 0 through 11) and adolescents (ages 12 through 17) in the household.
3. The AR is asked in the adult interview if either the AR or the ARSP is the parent or guardian for each child and Adolescent. Children and adolescents for whom the AR is the parent/guardian are associated with the AR and similarly for the ARSP.
4. If any adolescents are associated with the AR, then exactly one of these associated adolescents is randomly selected. Each associated adolescent has the same probability of selection in this step.
5. If any children are associated with the AR, then exactly one of these associated children is randomly selected. Each associated child has the same probability of selection in this step.

The last step is to compute the probability of selection for each sampled person. Since the adult was selected with equal probability, the probability of selection is just the inverse of the number of adults in a household.

The children and adolescents are sampled in two phases, so the probability of selection for a child or adolescent is the probability of selection of the adult multiplied by the conditional probability that the child or adolescent is selected given that the associated adult is selected. If the child or adolescent is associated with two adults (AR and ARSP), the probability of selection is the sum of the probabilities calculated in this way for each adult. In other words, the probability of sampling the person through the AR is computed and added to that the probability of sampling the person through the ARSP.

For example, consider the following hypothetical situation. A married couple has one child of their own (assigned to both the AR and the ARSP in Step 3) and there is one child who is not related to the ARSP but is the child of the AR. This child is associated with the AR (but not to the spouse of the AR). The within-household probability of sampling the child of both the AR and ARSP is 0.75. This is the sum of the probability of selecting the child via the AR (0.5 * 0.5) plus the probability of sampling the child via the ARSP (.5 * 1). The within-household probability of sampling the other child is 0.25, since the only way this child can be sampled is via the AR (0.5 * 0.5).

These probabilities are also discussed in Report 5: Weighting and Estimation. In that report, the inverse of the probability of selection is the initial weight for the adults, children, and adolescents.

5. ACHIEVED SAMPLE SIZES

This chapter summarizes the number of completed interviews in CHIS 2003 for the RDD strata and supplemental samples and the relationship between the targeted and the achieved numbers. As mentioned in the previous chapters, the targeted goals for CHIS 2003 were stated in terms of the total number of completed adult interviews obtained at the end of the data collection period. The actual number of completed interviews is a function of the number of telephones sampled, the within-household person sampling, and different reasons for attrition. These reasons are discussed in more detail in Chapter 3. Detailed information about the response rates is presented in *CHIS 2003 Methodology Series: Report 4 – Response Rates*.

5.1 Comparison to Goals

Table 5-1 gives the number of completed adult interviews by two methods of classifying the geographic area in which the sampled adult resides. The first column of completed interviews in the table uses the data on the county that was available at the time of sampling (and during the data collection period). As noted in Chapter 3 on sampling households, each telephone number is assigned to exactly one stratum for sampling purposes, but the number may actually be for a household in a different county. The third column in the table uses the self-reported residence county of the adult respondent. This classification is based on the geocoded location of the adult's residence derived from data collected on the county, ZIP code, address, and street intersection in the adult interview. It is the classification that is most appropriate for analysis of CHIS 2003 data. *CHIS 2003 Methodology Series: Report 3 – Data Processing Procedures* describes how the self-reported data were processed and how reporting discrepancies were resolved. The table gives the number of completed interviews as percentages of the targeted number of adult interviews set at the time of the design. The targeted goals by county for the RDD sample are given in Table 3-1. A percentage of 100 or greater indicates the targeted number of adult interviews was reached in the stratum.

Table 5-1 shows that CHIS 2003 surpassed the targets in all of the areas except for the Remainder of Alameda based on the sampling location information that was available at the time of data collection. For the self-reported location, 38 of the 45 areas surpassed the target number of completes, and 6 of the 7 strata that did not surpass the target were above 95 percent of the target. The only area that had a significant shortfall from the target goal was Hayward, where 788 adult interviews were completed and the target was 1,222 interviews. The discrepancies between the two location classifications are largely a function of how well the sampling classification matched with the self-reported classification. For smaller geographic areas like Hayward city, the sampling classification tends to be less precise but this varies by specific location.

Table 5-1. Number of completed adult interviews by sampling and self-reported stratum

	Sampling location		Self-reported location	
Area	Completed interviews	Percent of target	Completed interviews	Percent of target
State Total	42,044	105.1	42,044	105.1
Los Angeles	10,350	102.6	10,363	102.8
San Diego	2,310	101.4	2,319	101.8
Orange	2,231	104.1	2,186	102.0
Santa Clara	1,340	103.4	1,395	107.6
San Bernardino	1,238	102.2	1,244	102.7
Riverside	1,180	101.8	1,186	102.3
Alameda	4,734	118.7	4,647	116.5
Hayward	1629	133.4	788	64.5
Oakland	1975	130.3	1853	122.2
Remainder of Alameda	1,130	90.3	2,006	160.2
Sacramento	1,062	102.2	1,061	102.1
Contra Costa	820	102.5	897	112.1
Fresno	626	104.3	630	105.0
San Francisco	917	114.6	904	113.0
Ventura	617	102.8	630	105.0
San Mateo	609	101.5	596	99.3
Kern	537	107.4	549	109.8
San Joaquin	521	104.2	523	104.6
Sonoma	507	101.4	519	103.8
Stanislaus	549	109.8	531	106.2
Santa Barbara	504	100.8	497	99.4
Solano	510	102.0	503	100.6
Tulare	575	115.0	582	116.4
Santa Cruz	512	102.4	480	96.0
Marin	521	104.2	522	104.4
San Luis Obispo	503	100.6	506	101.2
Placer	507	101.4	513	102.6
Merced	520	104.0	537	107.4
Butte	564	112.8	567	113.4
Shasta	506	101.2	537	107.4
Yolo	517	103.4	514	102.8
El Dorado	503	100.6	506	101.2

Table 5-1. Number of completed adult interviews by sampling and self reported stratum (continued)

Area	Sampling location		Self-reported location	
	Completed interviews	Percent of target	Completed interviews	Percent of target
Imperial	529	105.8	528	105.6
Napa	505	101.0	513	102.6
Kings	531	106.2	528	105.6
Madera	512	102.4	506	101.2
Monterey, San Benito	520	104.0	542	108.4
Del Norte, Humboldt	529	105.8	525	105.0
Lassen, Modoc, Siskiyou, Trinity	419	104.8	423	105.8
Lake, Mendocino	409	102.3	396	99.0
Colusa, Glen, Tehama	425	106.3	397	99.3
Sutter, Yuba	460	115.0	451	112.8
Plumas, Nevada, Sierra	403	100.8	390	97.5
Alpine, Amador, Calaveras, Inyo, Mariposa, Mono, Tuolumne	412	103.0	401	100.3

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

Table 5-2 shows the number of completed child and adolescent interviews for the RDD sample. For these interviews, the targets were set overall rather than by county. The self-reported location classification is used in this table. The CHIS 2003 targeted goals were between 8,000 and 10,000 completed child interviews in the state and between 3,000 and 4,000 completed adolescent interviews in the state. In both cases, the achieved samples for children and adolescents were very close to the expected numbers.

Table 5-2. Number of completed child and adolescent completed interviews by self-reported location

Areas	Completed child interviews	Completed adolescent interviews
State Total	8,526	4,010
Los Angeles	2,112	925
San Diego	285	149
Orange	265	136
Santa Clara	2,112	925
San Bernardino	457	208
Riverside	466	201
Alameda	950	403
Hayward	356	158
Oakland	370	144
Remainder of Alameda	224	101
Sacramento	201	81
Contra Costa	163	87
Fresno	178	66
San Francisco	115	36
Ventura	127	59
San Mateo	110	54
Kern	124	64
San Joaquin	114	62
Sonoma	96	39
Stanislaus	119	65
Santa Barbara	107	59
Solano	113	67
Tulare	142	82
Santa Cruz	87	48
Marin	94	31
San Luis Obispo	83	46
Placer	97	56
Merced	141	69
Butte	98	53
Shasta	81	43
Yolo	102	56
El Dorado	94	55
Imperial	124	85
Napa	87	43
Kings	161	72
Madera	104	68
Monterey, San Benito	122	44

Table 5-2. Number of completed child and adolescent completed interviews by self-reported areas (continued)

Areas	Completed child interviews	Completed adolescent interviews
Del Norte, Humboldt	91	44
Lassen, Modoc, Siskiyou, Trinity	65	37
Lake, Mendocino	71	32
Colusa, Glen, Tehama	90	41
Sutter, Yuba	105	52
Plumas, Nevada, Sierra	53	38
Alpine, Amador, Calaveras, Inyo, Mariposa, Mono, Tuolumne	53	31

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

Table 5-3 shows the number of completed adult, child and adolescent interviews for the Korean and Vietnamese surname list samples. The supplemental sample targets were revised during the data collection period as experience was gained on the actual RDD sample yield. The second column of the table shows the revised target of the number of completed adult interviews. These targets were set for the State overall rather than by county. As with the RDD sample, the revised targets were generally achieved.

Table 5-3. Number of completed adult, child, and adolescent interviews by surname list sample

	Adult interviews			Completed child interviews	Completed adolescent interviews
Surname list sample	Revised target	Completed	Percentage of target		
Korean	49	112	228.6	24	6
Vietnamese	125	114	91.2	22	8

Source: UCLA Center for Health Policy Research, 2003 California Health Interview Survey.

REFERENCES

Anderson, J.E., Nelson, D.E., and Wilson, R.W. (1998). Telephone coverage and measurement of health risk indicators: Data from the National Health Interview Survey. *American Journal of Public Health,* 88, 1392-1395.

Blumberg, S., Luke, J., and Cynamon, M. (2004). Has cord-cutting cut into random-digit-dialed health surveys? The prevalence and impact of wireless substitution. *Proceedings of the Eighth Conference on Health Survey Research Methods,* Atlanta, GA.

Brick, J.M., and Waksberg, J. (1991). Avoiding sequential sampling with RDD. *Survey Methodology*, 17(1), 27-41.

Brick, J.M., Waksberg, J., Kulp, D., and Starer, A. (1995). Bias in list-assisted telephone surveys. *Public Opinion Quarterly*, 59(2) 218-235.

Casady, R., and Lepkowki, J. (1993). Stratified telephone survey designs. *Survey Methodology*, 19, 103-113.

Elliott, M R., Little R. J.A., and Lewitzky S. (2000). "Subsampling Callbacks to improve survey efficiency." *Journal of the American Statistical Association* 95:451 730-738.

Ford, E.S. (1998). Characteristics of survey participants with and without a telephone: Findings from the Third National Health and Nutrition Examination Survey. *Journal of Clinical Epidemiology,* 51, 55-60.

Giesbrecht, L.H., Kulp, D.W., and Starer, A.W. (1996). "Estimating coverage bias in RDD samples with Current Population Survey Data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 503-508.

Hansen, M.H., Hurwitz, W.N. & Madow, W.G. (1953). Sample Survey Methods & Theory, Jon Wiley & Sons, NY, Vols. I & II.

Kalton, G., and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society*, A, 149, 65-82.

Kish, L. (1949). "A procedure for objective respondent selection within the household", *Journal of the American Statistical Association* 44:380-87.

Kish, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, 8, 183-200.

Sudman, S., Sirken, M.G., and Cowan, C.D. (1988). Sampling rare and elusive populations. *Science*, 240,991-9

Tucker, C., Brick, J.M., Meekins, B., and Morganstein, D. (2004). Household telephone service and usage patterns in the U.S. in 2004. *Proceedings of the Survey Methods Section of the American Statistical Association* [CD-ROM].