

Chapter 3: Sample Design

By Adam Chu and Joseph D. Goldman

The primary goal of the sample design for the CSFII/DHKS 1994–96 was to obtain a nationally representative sample of noninstitutionalized persons residing in households in the United States for each of 40 analytic domains defined by sex, age (10 age groups), and income level (an "all-income" group and a "low-income" group) that met specified precision levels for estimates of mean day-1 saturated fat and iron intakes.¹ The specific precision goals required the coefficients of variation (*CV*'s) for mean saturated fat and iron intakes to be 3 percent or less for each of the 20 all-income sex-age domains and to be 5 percent or less for each of the 20 low-income sex-age domains. These precision goals were translated by Westat into 3-year sample size targets (table 2). In addition, the sample design specified that one day-1 intake respondent 20 years of age or older be selected for the DHKS from each household with at least one day-1 intake respondent age 20 or over. The design of the 3-year sample was such that the annual portions of the sample were roughly equal in size over the 40 analytic domains, and each year was nationally representative.

A complex, multistage, area probability sample design was used to select persons for the intake and DHKS interviews. The sample design was based on a Westat master sample that existed before the contract for the CSFII/DHKS 1994–96 was awarded.² The design included the selection of geographical areas called primary sampling units (PSU's), area segments within the sampled PSU's, households within the selected segments, and sample persons (SP's) within the households. The major features of the design are summarized below:

- The first-stage sample was a stratified sample of 62 PSU's consisting of metropolitan statistical areas (MSA's) or groups of counties. PSU's were selected within strata of approximately equal size, with probabilities proportional to the 1990 population.

1. For the CSFII/DHKS 1994–96, a single sample was selected that met precision requirements by income level. This differs from past CSFII/DHKS surveys where a separate sample of low-income persons was chosen in addition to the basic sample.

2. Persons living in group quarters or institutions, residing on military installations, and the homeless were excluded.

- Thirty-six area segments (consisting of census blocks or groups of blocks) were selected from each PSU, for a total of 2,232 area segments for the 3-year survey. The 36 segments selected from each PSU were divided into 12 sets of 3 segments each, and a set of 3 segments per PSU was assigned to each of the 12 quarters of the 3-year survey period.
- Within the sampled segments, lists of dwelling units (DU's) were prepared by Westat interviewers. More than 100,000 DU's were listed for each year of the survey. A self-weighting sample was selected from each listing. Approximately 9,500 DU's were selected for the first year, approximately 11,500 were selected for the second year, and approximately 12,000 were selected for the third year. The increased numbers of DU's selected did not necessarily result in increased numbers of SP's. Sampling rates also changed throughout the survey (see "Derivation of sampling rates and sampling messages" below).
- Within the occupied DU's identified during screening, households were identified and household members eligible for the survey were selected by a probability sampling process designed to achieve the specified sample sizes for various sex-age-income domains (see table 2).
- From households containing SP's 20 years of age or older who completed the day-1 intake interview, one SP was randomly selected for the DHKS.

Selection of Primary Sampling Units

At the first stage of sampling, the entire United States was divided into PSU's consisting of MSA's, counties, or groups of counties. The sampling frame of PSU's was created from county-level data contained in the 1990 Census Public Law 94-171 (Public Law 94) and data files from the Bureau of Economic Analysis, U.S. Department of Commerce (U.S. Department of Commerce-Bureau of the Census 1991a). The Public Law 94 data file provided county-level population counts by race and Hispanic origin, while the Bureau of Economic Analysis file provided the corresponding income information.

Because of their size, the New York MSA was divided into three PSU's and the Los Angeles and Chicago MSA's were each divided into two PSU's. Each of the other MSA's constituted a single PSU. Counties outside MSA's were grouped, as necessary, to form PSU's that (1) had a minimum 1990 population of 15,000 people, (2) were as internally heterogeneous as possible, and (3) were still small enough to permit convenient travel across the PSU by interviewers. From the

more than 3,000 counties in the United States, a total of 1,404 PSU's was created, and 62 PSU's were selected for use in the CSFII/DHKS 1994–96.

The 24 PSU's with the largest populations were included with certainty. The remaining (noncertainty) PSU's were then assigned to 1 of 38 strata of approximately equal size (in terms of 1990 population), and one PSU was selected from each stratum with probability proportional to the 1990 population. Stratification factors used to select the noncertainty PSU's included the region of the country (four census regions), whether or not the PSU was an MSA and the population size of the MSA, percentage of the population that was black or Hispanic, and per capita income. Among the noncertainty strata, 26 were MSA strata and 12 were non-MSA strata. The distribution of the sampled PSU's by census region and MSA status is summarized in table 3. The nature of the PSU's does not allow for state-level estimates.

Selection of Area Segments

The second-stage sampling units were area segments, which were defined to be individual census blocks or a group of blocks. A sample of 36 area segments was randomly selected from each PSU, with probability proportional to population. The 36 segments were then divided into 12 sets of 3 segments each, and a set of 3 segments per PSU was assigned to each of the 12 quarters of the 3-year survey period. Segments were assigned to the quarters of the year in a balanced, random manner to ensure a wide spread of the segment sample within each quarter for each PSU. This balanced sampling was carried out to improve sampling precision by reducing the design effects resulting from the homogeneity of persons within segments. This method also achieved the general sample design requirement of having data collection spread evenly over the 3 years of the survey and over the quarters of the year.

As part of the sampling process, a frame of area segments for each of the 62 sample PSU's was created. This frame was constructed from the Census Bureau's 1990 Public Law 94 datatape, which contains population, housing counts, and limited geographic information for each block in the United States (U.S. Department of Commerce–Bureau of the Census 1991a). To ensure that the segments would be of sufficient size for use in sampling, small blocks were combined with adjacent blocks to form segments that had a minimum expected size of 60 DU's. After the frame was constructed, the area segments were sorted before sample selection into minority strata (based on black and Hispanic households) and geographically within minority strata. For each of the 3 years of

the study, a systematic sample of 12 area segments was selected from the sorted frame, with probabilities proportional to the number of DU's in the segment.

A national sample of segments that Westat had previously selected and listed in the selected PSU's was used to reduce sampling costs. The sample developed for the National Adult Literacy Survey (NALS) used basically the same sampling procedures required for the CSFII/DHKS 1994–96, except that high-density minority segments were selected at about twice the rate of the nonminority segments. For the NALS, the segments were deliberately made much larger than needed so that they could serve as the equivalent of a master segment sample that could be used for other studies. Fifty-six percent of the 2,232 segments required for the CSFII/DHKS sample could be drawn from the previously selected NALS segments. The remaining segments were selected to yield the desired overall probabilities of selection, while maximizing the overlap with the NALS sample. NALS listings were updated through standard quality control procedures.

Selection of Dwelling Units

The sample of DU's was selected from the sample of area segments. The procedures used to select the DU's included the creation of segment-level lists of DU's, use of special procedures for handling a few extremely large segments (chunks) in the listing process, the selection of DU's from the segment listings, and special field procedures used to verify and update the listing information.

The purpose of listing was to create a list of DU's from which a sample could be selected for interviewing. For the sample to be representative of the population of interest, it was essential that the listing be carried out accurately and systematically, so that every DU in a designated segment was included. The process of listing involved an interviewer walking or driving through every street, road, alley, or boundary in the segment and recording on forms the address and description of every DU within the boundaries of the selected segments. The maps necessary to list the segments were generated using the U.S. Census Bureau's map-producing database called TIGER (Topologically Integrated Geographic Encoding and Referencing) (U.S. Department of Commerce–Bureau of the Census 1991c). The TIGER file is a geographic database where all map features are digitized and stored along with attribute information.

Census data indicated that some of the sampled segments were very large. To reduce the listing workload in the large segments, an additional stage of sampling was introduced. In general, these segments (defined as segments with an estimated 500 or more DU's) were divided into two or more smaller chunks of

approximately equal size, and one chunk was selected for listing with probability proportional to estimated size. Of the 744 segments (including NALS segments) selected for each year of the CSFII/DHKS, 54 were chunked using these procedures in 1994, 38 were chunked in 1995, and 58 were chunked in 1996. Although the selected chunks were treated like all other segments in the subsequent stages of selection, their probabilities of selection were properly adjusted to reflect the additional stage of selection.

For the first year, a sample of about 9,500 DU's was selected from the 744 segments (or chunks) designated for the first year of data collection. It was estimated that 9,500 DU's were necessary to yield approximately one-third of the required number of SP's within each of the sex-age-income groups defined by the survey design specifications prior to data collection. This estimate took into account the percentages of individuals in each sex-age group living in households, the percentages of individuals in each sex-age group living in households at or below 130 percent of the Federal poverty guidelines (U.S. Department of Health and Human Services 1994), projected response rates, a projected rate for vacant DU's, and a safety factor allowing for random sampling variation. The number of DU's selected was increased to about 11,500 for the second year and about 12,000 for the third year because the sampling rates of individuals changed throughout the survey (see "Derivation of sampling rates and sampling messages" below). The increased numbers of DU's selected did not necessarily result in increased numbers of SP's overall. The procedure for selecting DU's for the first year follows. The same procedure was used for the second and third years. The only change was the number of DU's selected.

To select the sample, the overall national sampling rate (f) was computed by dividing 9,500 by the estimated number of DU's (\hat{N}) based on the DU counts obtained during listing. Specifically, \hat{N} was calculated from the formula:

$$\hat{N} = \sum_{h=1}^{62} \left[\frac{1}{P_h} \right] \sum_{j=1}^{12} \left[\frac{N_{hj}^L}{P_{hj}} \right] \quad [1]$$

where

P_h is the probability of selecting PSU h,
 N_{hj}^L is the number of DU's listed in segment j in PSU h, and
 P_{hj} is the within-PSU probability of selecting segment j in PSU h for the first year of the survey.

For the NALS segments, N_{hj}^L reflected the numbers of DU's originally listed for the NALS (not including any new or missed structures added through the "missed structure" or "missed DU" procedures described below). This is because the selection of DU's was restricted to those DU's originally listed for NALS. However, it does not mean that new construction had no chance of selection from NALS segments. New construction (and also DU's that were missed in the original NALS listing process) still had appropriate chances of selection through the missed structure and missed DU procedures. As documented below, the within-segment sampling rates used to select the DU's were designed to produce a self-weighting national sample of approximately 9,500 DU's. Note that the procedure for selecting the DU's within the NALS and non-NALS segments was slightly different because of the desire to avoid selecting those DU's previously selected for NALS in the NALS segments.

Selection of dwelling units outside of National Adult Literacy Survey segments

N_{hj}^L denotes the number of DU's that were listed in non-National Adult Literacy Survey (non-NALS) segment j in PSU h. The N_{hj}^L DU's in the segment were then subsampled with equal probabilities at a rate of

$$f_{hj}^{(w)} = \frac{f}{P_h P_{hj}} \quad [2]$$

where

P_h is the probability of selecting the PSU and
 P_{hj} is the conditional probability of selecting the segment within the PSU.

The within-segment sampling rate, $f_{hj}^{(w)}$ given by formula 2 was designed to give each DU in the segment an overall probability of selection equal to f (that is, $P_h P_{hj} f_{hj}^{(w)} = f$). The actual selection of DU's within a segment was accomplished by first creating a file of unique line numbers corresponding to the DU's listed in the segment, and then selecting the line numbers systematically using a random start and a skip interval equal to $1 / f_{hj}^{(w)}$. A systematic sampling algorithm was used to make the selections (Hansen et al. 1953).

Selection of dwelling units in National Adult Literacy Survey segments

Let N_{hj}^L denote the number of DU's that were originally listed for the NALS in segment j in PSU h . For NALS segments, the count N_{hj}^L does not include any structures or DU's that were added as a result of the missed structure or missed DU procedures. Of the N_{hj}^L DU's in the segment that were originally listed for NALS, the n_{hj}^{NALS} DU's sampled for the NALS were identified and excluded from the sampling process. The remaining $N_{hj}^L - n_{hj}^{NALS}$ DU's were then subsampled at a rate of

$$f_{hj}^{(w)} = \frac{f}{P_h P_{hj} \left[\frac{N_{hj}^L - n_{hj}^{NALS}}{N_{HJ}^L} \right]} \quad [3]$$

where the term $N_{hj}^L - n_{hj}^{NALS} / N_{hj}^L$ in the denominator of formula 3 is the probability that a DU in the segment was not previously selected for NALS.

The selection of DU's within a NALS segment was accomplished by first creating a file of unique line numbers corresponding to the DU's listed in the segment, deleting the line numbers corresponding to the DU's previously selected for NALS, and then systematically selecting the line numbers using a random start and a skip interval equal to $1 / f_{hj}^{(w)}$. The overall sampling rate for DU's in the NALS segments is the same as that in the non-NALS segments (that is, $P_h P_{hj} (N_{hj}^L - n_{hj}^{NALS} / N_{hj}^L) / N_{hj}^{NALS} f_{hj}^{(w)} = f$).

Missed structure and dwelling unit procedures

Two separate quality control procedures were used to verify and update the listing information for all of the segments selected for the CSFII/DHKS 1994–96. Both procedures were conducted during data collection. The first of these, referred to as the missed structure procedure, was applied whenever the first DU in the

segment was selected for the CSFII/DHKS sample. Two versions of the missed structure procedure were used. The original procedure was used during the first 2 years of the survey. The procedure was modified at the start of the third year of the survey to account for the possibility that large amounts of new construction may have occurred in the NALS segments. Descriptions of both versions follow.

When a segment was designated for the missed structure procedure during the first 2 years of the survey, the interviewer recanvassed the entire segment, and all DU's not previously listed were added to the sample (except as noted below). Because the probability of selecting the first DU in a segment was equal to the within-segment sampling rate, all of the added DU's were selected at the same overall rate (that is, they had the same overall probability of selection) as the rest of the sample.

The above rule for designating the missed structure procedure segments during the first 2 years of the survey applied to the NALS and non-NALS segments. Because DU's selected for the NALS were excluded from the CSFII/DHKS sample, NALS segments that were designated for the missed structure procedure in the NALS were not designated for this procedure in the CSFII/DHKS. In effect, the updating work performed for the NALS was ignored for the CSFII/DHKS. However, no bias was introduced because the new or missed DU's still had their appropriate chances of selection for the CSFII/DHKS.

For the third year of CSFII/DHKS 1994–96, the missed structure procedure was modified so that the rules for designating the missed structure segments were different depending on whether the segment was a NALS or a non-NALS segment. For the non-NALS segments, the original rules applied; that is, a non-NALS segment was designated for the missed structure procedure if the first DU in the segment was selected for the sample. However, for the NALS segments, a modified rule was adopted. The modified rule was designed because large amounts of new construction could have occurred since the NALS segments were originally listed in 1991. Under the modified rule, an NALS segment was designated for the missed structure procedure if any of the first four DU's in the segment were selected for the CSFII. Therefore, on average, the NALS segments were designated for the missed structure procedure at four times the rate of non-NALS segments.

In those segments selected for the missed structure procedure, the interviewer prepared a list of all DU's that were not included in the original listing forms (that is, the new or missed DU's). This information was then sent to the central office, where a subsample of the new or missed DU's was selected by computer at rates

designed to yield the same overall probabilities of selection as the other DU's in the sample. Therefore, in general, all of the new or missed DU's in the non-NALS segments were added to the CSFII sample. On the other hand, only one in four of the new or missed DU's in the NALS segments were added to the sample to compensate for the fact that the NALS segments had four times as many chances of being selected for the missed structure procedure.

The second procedure, referred to as the missed DU procedure, applied to structures containing many DU's (for example, apartment buildings) and all DU's listed at a single address. If the first DU in the given structure was selected for the CSFII/DHKS sample, then the entire structure was checked to identify DU's that may have been omitted from the listing sheets. Any missed DU's found by this process were added to the sample.

To keep the interviewing workload to manageable levels within the segment, maximum limits were established for the number of missed or new DU's that could be added to the sample. These limits were 10 per segment for the missed structure procedure and 4 per structure for the missed DU procedure. When the actual numbers of missed DU's exceeded these limits, a subsample of the missed DU's was retained in the sample. In the first year of the CSFII/DHKS, subsampling was required for 4 of the 93 segments when the missed structure procedure was applied. In the second year, subsampling was required for 6 of the 126, and in the third year, subsampling was required for 8 of the 282 segments when the missed structure procedure was applied. Subsampling of the missed DU's was not required for any structures when the missed DU procedure was applied in any year. During the first year of the survey, 77 DU's were added to the sample through the missed DU procedure and 128 DU's were added through the missed structure procedure. During the second year, 100 DU's were added to the sample through the missed DU procedure and 219 DU's were added through the missed structure procedure. During the third year, 96 DU's were added to the sample through the missed DU procedure and 464 DU's were added through the missed structure procedure.

Results of the dwelling unit sampling process

A total of 32,932 DU's was selected for the 3 years of the survey. In addition, 1,084 DU's were added in the field as a result of the missed structure procedure and the missed DU procedure. Therefore, the total number of DU's included in the sample was 34,016. Of these, 33,560 completed either the full or abbreviated screener questionnaire or were vacant or non-DU's, and 4,189 were either vacant or non-DU structures. Of the 29,827 occupied DU's, 9,664 (32 percent) had

household members who were eligible for the survey. The percentage of screened households with eligible SP's decreased over 3 years, from 39 percent to 33 percent and then to 28 percent, as a result of the changes in the sampling rates of individuals. The results of the DU sampling process are summarized in table 4.

Selection of Sample Persons for Intake Interviews

The CSFII 1994–96 was designed to obtain a sample that would produce estimates with equivalent precision over the sex-age domains, for both the total population and the low-income population. To obtain the targeted numbers of individuals, different sex-age domains were sampled at different rates. The approach used to select persons for the intake interviews was to designate subsets of households where only persons meeting specified sex-age/income criteria would be included in the sample. For example, for one predesignated subset of households in the DU sample, only children between the ages of 1 and 2 years and low-income males between the ages of 50 and 59 years were to be included in the sample. Sampled households were randomly assigned to the various subsets to ensure the unbiased selection of SP's for the study. In addition, all infants under 1 year of age in households that contained at least one SP 1 year or older were included in the sample.

To facilitate the selection of SP's in the field, each screening questionnaire carried a sampling message specifying the characteristics of the persons to be included. These sampling messages were assigned at Westat's home office and the interviewers had no discretion as to whom to include. A total of 24 distinct sampling messages were employed for the first year of the CSFII/DHKS 1994–96—21 messages were employed in the first half of the second year, a slightly different set of 21 messages were employed in the second half of the second year, 13 messages were employed in the first half of the third year, and 17 messages were employed in the second half of the third year.

The proportion of households that received a particular message was determined to satisfy the target sampling rates for the various sex-age-income domains. The number and configuration of the sampling messages was a function of these sampling rates. The initial 24 messages used in the first year of CSFII/DHKS 1994–96 were derived from estimates based on a previous survey and on the pilot study experience. Once screening operations began and the results could be analyzed, the target sampling rates were adjusted to meet the sex-age-income domain goals as closely as possible. New sets of sampling messages were introduced at the beginning and midway into the final 2 years of the survey.

After completing the listing of household members, the interviewer identified which, if any, of the household members were eligible to be interviewed. A total of 19,830 SP's were identified through the screening process during the 3 years of the survey, with 6,868 in the first year, 6,576 in the second year, and 6,386 in the third year.

Derivation of sampling rates and sampling messages

Year 1. The form of the sampling messages used in the first year of CSFII/DHKS 1994–96 to select SP's was determined as follows. First, estimates of the number of persons in each sex-age-income domain were obtained from the March 1992 Current Population Survey (CPS) public use file (U.S. Department of Commerce–Bureau of the Census 1993). Second, coverage rates from the 1992 National Health Interview Survey (NHIS) (U.S. Department of Health and Human Services–National Center for Health Statistics 1994) were applied to the March 1992 CPS counts to obtain estimates of the numbers of persons who would be covered by an area probability sample. CPS estimates included adjustments to compensate for the known undercounting of certain groups of individuals and were expected to be somewhat larger than the corresponding counts obtained from the CSFII/DHKS listing operations where similar undercounting could be expected. Without the downward adjustment of the CPS estimates through application of NHIS coverage rates, the derived sampling rates might have been underestimated. Initial sampling rates were then defined for each sex-age-income group as the ratio of the sample size targets to the downward-adjusted, estimated population counts.

Third, some adjustments to the initial sampling rates were implemented. For 5 of the 20 sex-age groups, the proportion of low-income persons was high enough so that using the initial sampling rate for the total population would achieve both the all-income sample size target and the low-income sample size target. For these groups, the low-income sampling rates were adjusted by replacing them with the all-income sampling rate.

For the remaining 15 groups, different all-income and low-income sampling rates were used. The initial low-income rates were retained without adjustment. The all-income rates were adjusted by replacing them with the rates expected to obtain $n^{\text{non-L}} = n^{\text{all}} - n^{\text{L}}$ non-low-income sample persons from the all-income population, where n^{all} and n^{L} are the all-income and low-income sample size targets. Both the all-income and low-income sample size targets were expected to be met as a result of this adjustment to the initial rates and the combination of the sampling of the all-income population and a supplemental sampling of the low-income population.

Table 5 shows the sample size targets, estimated population counts, and initial and adjusted sampling rates for each sex-age-income group. Column 4 shows the CSFII/DHKS sample size targets from table 2. As stated above, the initial sampling rates are the ratio of the sample target sizes in column 4 and the population counts in column 5.

Once the adjusted sampling rates were calculated for each sex-age-income group, the groups were ordered by the magnitude of the rates and, in some cases, combined with other groups with similar sampling rates. Where groups were combined, the highest sampling rate among the groups was assigned to each of the groups in the combination. The result was 24 distinct groups, each consisting of 1 or more of the 40 sex-age-income groups. Table 6 shows these combined sex-age-income groups and their adjusted and final sampling rates.

Table 7 shows the 24 sampling messages. The messages are cumulative. For example, message 1 indicates that all children age 1 and 2 and low-income males age 50–59 would be selected from a household assigned that message, while those persons and low-income males age 60–69 would be selected from a household assigned message 2. Additionally, all infants under 1 year of age were selected only if another person 1 year of age or older was also selected through the sampling messages. The rightmost column of table 7 shows the proportion of all DU's selected for the sample assigned each sampling message. That is, 16.63 percent of all DU's were assigned message 1 and 17.35 percent of all households were assigned message 24. The proportion of DU's assigned to sampling message i was calculated from the formula:

$$prop_i = \frac{(r_i - r_{i+1})}{r_1} \quad [4]$$

where

- r_i is the corresponding final sampling rate given in the last column of table 5,
- r_{i+1} is the final sampling rate given in the preceding row of the table (where $r_{25} = 0$ by definition), and
- $r_1 = 0.2004$ is the targeted sampling rate (corresponding to the last row of the table).

The sampled DU's within each PSU were randomly assigned to the various messages in the proportions given in the rightmost column of table 7. This was accomplished by computing $N_{DU}prop_i$ (rounded to the nearest integer) for each

message $i = 1, 2, \dots, 24$, where N_{DU} is the number of sampled DU's in the PSU, and then randomly assigning the required number of DU's to message i .

Year 2. The numbers of completed day-1 intakes obtained during the first year of the survey generally met or exceeded the designated 1-year targets, with some exceptions. Shortfalls occurred in four all-income domains and in seven low-income domains. To compensate, the sampling rates established for the first year of the survey were modified to make up for the shortfall equally in the subsequent 2 years of the survey. For example, suppose that, at the end of the first year, 60 day-1 intake interviews were obtained for a domain where the 3-year target was 207. At that rate, 180 completed interviews would be available at the end of the survey, short of the target of 207. To make up for the shortfall, the original sampling rate was increased by about 20 percent to obtain an expected 74 completed interviews in each of the next 2 years of the survey $(207 - 60) / 2 = 74$. Similarly, for those domains where there was an excess of completed interviews in the first year, a corresponding downward adjustment was made to the original sampling rates.

Ideally, it would have been desirable to use all of the information available at the end of the first year to make the necessary changes for the second year. Unfortunately, this was not possible because of the amount of time needed to process the survey results and to prepare interviewer materials for the first quarter of the second year of the survey. In order to proceed with the preparation of materials the sample yield results from only the first two quarters of the first year were used to design the sampling rates for the second year.

The procedures used to construct the sampling messages for the second year were analogous to those previously described for the first year. Once the initial sampling rates were calculated for each sex-age-income group, the groups were ordered by the magnitude of the rates, and, in some cases, combined with other groups with similar rates. The result was 21 distinct groups.

Year 3. The adjustments in sampling rates were successful in eliminating 7 of the 11 shortfalls in sample yields observed after the first year of data collection. However, the sample yields of several other groups decreased and at the end of the second year, there were 10 domains where the 3-year sample size targets would probably not be met if rates were not adjusted. These shortfalls occurred in four all-income domains and six low-income domains. To compensate for these shortfalls, the sampling rates established for the second year of the survey were modified to make up for the shortfall in the final year of the survey.

As was done at the end of the first year to prepare for the second year, it would have been desirable to use all of the information available at the end of the second year of the survey to make the necessary changes for the third year. This was not possible, however, and in order to proceed with the preparation of materials, the sample yield results from the first three-quarters of the second year, along with complete results from the first year, were used to design the sampling rates for the third year. A different approach than that used for the second year's adjustments was taken. The sampling rates for the third year were constructed by first projecting the sample yields for each of the sex-age-income domains through the first 2 years. These counts had to be projected because preparation for the third year had to be completed before the second year data collection was completed. It was necessary to project the sample yields for the fourth quarter of year 2 for the all-income domains and for the third and fourth quarters for the low-income domains. Once completed, the difference between the 3-year target, and the projected 2-year sample yields provided a new target for the third year and sampling rates and sample messages were designed to meet those targets. For example, for all-income males age 1–2, the actual yield for the first year was 255 and the actual yield for the first three-quarters of the second year was 204. A projection of 524 males was calculated from these actual counts and an estimate of the last quarter's yield. The 3-year target for all-income males age 1–2 was 719, so the target for the third year was calculated as $(719 - 524 = 195)$. A yield ratio, the ratio of the number of completed day-1 intake interviews to the number of DU's where an all-income male age 1–2 was assigned for sampling, was calculated from the results of the first seven quarters. The projected number of DU's needed for sampling to meet the target for males age 1–2 was then calculated by dividing the third year target by the yield ratio. In this example, the yield ratio was 0.02687, so the number of DU's expected in the required 195 day-1 interviews was calculated as $195 / 0.02687 = 7,256$.

The procedures used to construct the sampling messages for the third year were analogous to those previously described for the first year. Once the initial numbers of DU's to be sampled were calculated for each sex-age-income group, the groups were ordered by the magnitude of these numbers, and, in some cases, combined with other groups with similar requirements. The result was 13 distinct groups.

As in the second year, adjustments were made to the sampling messages midway through the third year. Unlike the second year, where the changes were made to the existing set of messages, the mid-third-year adjustments were made by recomputing the required rates to reflect actual returns through the first quarter of the third year and then reworking the sampling messages using these rates.

Classification of households to income classes

Under the procedures adopted for the CSFII/DHKS, the screener (see chapter 4) contained a question on income status (Q S14) that was asked only when necessary during screening because of the belief that asking about income during the initial contact might increase nonresponse to the survey. Therefore, if the sampling message indicated that income information was unnecessary, the question was not asked. For example, one message indicated that all persons (1 year of age or older) in these households were to be included in the sample regardless of income level. Similarly, another message selected persons 1–2 years of age and low-income males 50–59 years of age. If a household assigned this message did not include males 50–59 years of age, the sampling of SP's could proceed without collecting income data in the screener. In these cases, the income information was requested during the household interview, using an identically worded question (Q H47a) from the more detailed household questionnaire.

Occasionally, the interviewers were unable to obtain the income information necessary to select SP's for the intake interviews. In such cases, a rule based on the composition of the household was used to assign the household to one of the income groups for sampling purposes. The rule used was the following: If the household contained one or more children under 6 years of age, but no males 18 years of age or over, it was treated as low income for sampling purposes. Otherwise, the household was treated as non-low income. This rule was expected to be reasonably effective in identifying low-income households because more than 60 percent of children under 6 years of age who live in households headed by a female with related children under 6 years and no spouse present are living below Federal poverty guidelines (U.S. Department of Commerce–Bureau of the Census 1991b).

It should be noted that the sampling rule given above was adopted simply to facilitate the sampling of SP's in the field. Some households that were classified as low income by this rule may have turned out to be non-low income, and vice versa. For base weighting purposes, such households were weighted according to their income status as determined by the sampling rule and not their actual income status. However, for the purpose of determining sample yields, the response to either Q S14 of the screener questionnaire or Q H47a of the household questionnaire was generally used to establish income status. Where a response to Q S14 or Q H47a was not available, a series of five income-status imputation rules were used to determine low-income status for the purpose of determining sample yields. The five rules were applied sequentially, that is, if rule 1 could not be used, then rule 2 was used, and so on. The five income-status imputation rules were:

1. Annual income from household questionnaire items H52 or H53 was used, along with household size, to determine low-income status.
2. Monthly income from household questionnaire items H57a–H57f was totaled and used, along with household size, if rule 1 could not be applied.
3. Household questionnaire item H58 used a handcard to ask if last month’s income was above or below the appropriate low-income cutoff based on household size. The result of this question was used if the previous two rules could not be applied.
4. Household questionnaire item H59 asked about food stamp use. If the answer was yes, the household was assigned to the low-income group. If the answer was no, the household was assigned to the non-low-income group. This rule was used only if the previous three rules could not be applied.
5. Finally, if none of the above four rules could be applied, the sampling rule, based on household composition was used. Under the sampling rule, households with one or more children under 6 years of age and no males 18 years or older were treated as low-income. All other households were treated as non-low-income.

It was necessary to use these rules to classify about 2 percent of all SP’s with day-1 intakes as either low-income or non-low-income.

Results of SP sampling process

Table 8 summarizes the number of SP’s eligible for intake interviews, the corresponding numbers completing the first intake interview, and the success of the survey process in achieving the sample size goals. As shown, the sample size goal was met or exceeded for 14 of the 20 all-income sex-age domains. For all of the remaining six all-income sex-age domains, at least 98 percent of the CSFII goals were achieved. Among the low-income domains, the sample size goals were met or exceeded for 14 of the 20 sex-age domains. For four of the remaining six low-income sex-age domains, at least 96 percent of the CSFII target was achieved. The two low-income domains with the greatest shortfalls were females 50 to 59 years of age (about 9 percent short of the goal) and males 40 to 49 years of age (about 6 percent short of the goal).

Selection of Sample Persons for the DHKS 1994–96

Respondents for the DHKS were selected from among SP's 20 years of age and over who had completed the day-1 intake interview without a proxy.³ Only one DHKS respondent per household was selected in households with eligible participants. In households with more than one CSFII participant 20 years of age or over, one of the participants was selected randomly in the field using a specially designed sampling program in each interviewer's laptop computer. Unlike the intake interviews, there were no specific numerical sample size targets for the DHKS. However, there was the requirement that the distribution of DHKS respondents by age, sex, and income be similar to that of the corresponding intake respondents. Although it was recognized that restricting the DHKS sample to only one respondent per household might distort the distribution of DHKS respondents somewhat, the random sampling procedures used to select respondents were reasonably effective in meeting the study goals. As table 9 shows, the distribution of SP's selected for the DHKS and the corresponding distribution of DHKS respondents are generally comparable to the distribution of SP's completing the day-1 intake interview. The selection was made with probability assigned to maintain distributions of all-income and low-income individuals in the 6 sex-age groups age 20 years and over in the DHKS that conformed approximately to the corresponding distributions of individuals in the CSFII. Approximately one-half of the households had more than one eligible SP for the DHKS. In all 3 years of the survey, 6,294 individuals were selected into the DHKS 1994–96 sample, 2,047 in the first year, 2,159 in the second year, and 2,088 in the third year.

3. In 1994–96, 191 SP's age 20 or older completed the day-1 intake with the assistance of a proxy.

References

Hansen, M., W. Hurwitz, and W. Madow. 1953. *Sample survey methods and theory*. Vol. 1. John Wiley & Sons, New York.

U.S. Department of Commerce, Bureau of the Census. 1991a. *Census of population and housing, 1990: Public law 94-171 data*. Machine-readable data file.

U.S. Department of Commerce, Bureau of the Census. 1991b. *Poverty in the United States: 1990*. Current Population Reports, Series P-60, No. 175.

U.S. Department of Commerce, Bureau of the Census. 1991c. *TIGER/Line census files, 1990*. Machine-readable data file.

U.S. Department of Commerce, Bureau of the Census. 1993. *Current population survey, March 1992*. Machine-readable data file.

U.S. Department of Health and Human Services. 1994. *Poverty guidelines*. Federal Register 59(28):6277.

U.S. Department of Health and Human Services, National Center for Health Statistics. 1994. *1992 national health interview survey*. CD-ROM Series 10, No. 6, SETS Version 1.21. U.S. Government Printing Office, Washington, DC.

Table 2. Sample size targets, CSFII/DHKS 1994–96

Sex and age (years)	Sample size targets	
	Low income*	All income (total sample)
Male		
1–2	207	719
3–5	207	719
6–11	207	719
12–19	207	719
20–29	207	793
30–39	207	850
40–49	207	850
50–59	207	850
60–69	207	850
70 and over	207	793
Female		
1–2	207	719
3–5	207	719
6–11	207	719
12–19	207	719
20–29	207	739
30–39	207	793
40–49	207	850
50–59	207	850
60–69	207	793
70 and over	207	719
Total	4,140	15,482

* The income level used during the screening process corresponded to 130 percent of the Federal poverty guidelines (U.S. Department of Health and Human Services 1994), which are based on household size and income. This income level was selected because it is the same as one of the income criteria used to determine whether nonelderly households are eligible to participate in the Food Stamp Program. Not all households meeting the criteria are eligible for food stamps; other criteria, such as asset limitations, must also be met. The CSFII 1994–96 screened households for income level only, not for food stamp eligibility.

Table 3. Distribution of PSU's by census region and MSA status, CSFII/DHKS 1994-96

Census region	Type of PSU			Total
	Certainty MSA	Noncertainty MSA	Non-MSA	
Northeast	6	6	1	13
Midwest	5	8	4	17
South	6	7	5	18
West	7	5	2	14
Total	24	26	12	62

Table 4. Results of the DU sampling process, CSFII 1994–96

Survey year	DU's selected from listings	DU's added in the field	Total DU's in sample	Vacant or non-DU's	Occupied DU's with eligible SP's *	Occupied DU's with no eligible SP's	Non-responding DU's**
1994	9,423	205	9,628	1,161	3,266	5,067	134
1995	11,504	319	11,823	1,337	3,379	6,954	153
1996	12,005	560	12,565	1,691	3,019	7,686	169
1994–96	32,932	1,084	34,016	4,189	9,664	19,707	456

* Eligible SP's refers to household members designated for intake interviews by the SP sampling process.

** Nonresponding DU's are those where a screener questionnaire was not completed.

Table 5. Sample size targets, estimated population counts, and initial and adjusted sampling rates, CSFII/DHKS 1994

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sex	Age (years)	Income	Sample size target (3 years)	Population counts based on CPS totals and NHIS coverage (× 1,000)	Initial sampling rate (× 1,000)	Sample for sex-age group meets target for low-income(*)	Adjusted sampling rate (× 1,000)
Male	1-2	All	719	3,612	0.1991	*	0.1991
Male	3-5	All	719	5,248	0.1370	*	0.1370
Male	6-11	All	719	10,627	0.0677		0.0651
Male	12-19	All	719	12,682	0.0567		0.0514
Male	20-29	All	793	16,189	0.0490		0.0421
Male	30-39	All	850	18,454	0.0461		0.0397
Male	40-49	All	850	14,296	0.0595		0.0496
Male	50-59	All	850	9,844	0.0863		0.0730
Male	60-69	All	850	8,844	0.0961		0.0845
Male	70+	All	793	7,559	0.1049		0.0965
Female	1-2	All	719	3,587	0.2004	*	0.2004
Female	3-5	All	719	5,273	0.1363	*	0.1363
Female	6-11	All	719	10,240	0.0702		0.0694
Female	12-19	All	719	12,530	0.0574		0.0533
Female	20-29	All	739	16,474	0.0449		0.0408
Female	30-39	All	793	20,325	0.0390		0.0348
Female	40-49	All	850	16,113	0.0528		0.0452
Female	50-59	All	850	10,927	0.0778		0.0685
Female	60-69	All	793	10,554	0.0751		0.0706
Female	70+	All	719	11,056	0.0650	*	0.0650
Male	1-2	Low	207	1,184	0.1749		0.1991
Male	3-5	Low	207	1,517	0.1364		0.1370
Male	6-11	Low	207	2,758	0.0750		0.0750
Male	12-19	Low	207	2,726	0.0759		0.0759
Male	20-29	Low	207	2,264	0.0915		0.0915
Male	30-39	Low	207	2,249	0.0920		0.0920
Male	40-49	Low	207	1,322	0.1565		0.1565
Male	50-59	Low	207	1,034	0.2002		0.2002
Male	60-69	Low	207	1,239	0.1671		0.1671
Male	70+	Low	207	1,487	0.1392		0.1392
Female	1-2	Low	207	1,118	0.1852		0.2004
Female	3-5	Low	207	1,552	0.1334		0.1363
Female	6-11	Low	207	2,863	0.0723		0.0723
Female	12-19	Low	207	2,921	0.0709		0.0709
Female	20-29	Low	207	3,446	0.0601		0.0601
Female	30-39	Low	207	3,471	0.0596		0.0596
Female	40-49	Low	207	1,884	0.1099		0.1099
Female	50-59	Low	207	1,542	0.1342		0.1342
Female	60-69	Low	207	2,251	0.0920		0.0920
Female	70+	Low	207	3,883	0.0533		0.0650

Table 6. Final sampling rates assigned to each message, CSFII/DHKS 1994

(1)	(2)	(3)	(4)	(5)	(6)
Sex	Age (years)	Income	Adjusted sampling rate (× 1,000)	Sampling message number	Final sampling rate (× 1,000)
Female	30–39	All	0.0348	24	0.0348
Male	30–39	All	0.0397	23	0.0397
Female	20–29	All	0.0408	22	0.0408
Male	20–29	All	0.0421	21	0.0421
Female	40–49	All	0.0452	20	0.0452
Male	40–49	All	0.0496	19	0.0496
Male	12–19	All	0.0514	18	0.0514
Female	12–19	All	0.0533	17	0.0533
Female	30–39	Low	0.0596	16	0.0601
Female	20–29	Low	0.0601	16	0.0601
Female	70+	All	0.0650	15	0.0651
Female	70+	Low	0.0650	15	0.0651
Male	6–11	All	0.0651	15	0.0651
Female	50–59	All	0.0685	14	0.0694
Female	6–11	All	0.0694	14	0.0694
Female	60–69	All	0.0706	13	0.0709
Female	12–19	Low	0.0709	13	0.0709
Female	6–11	Low	0.0723	12	0.0730
Male	50–59	All	0.0730	12	0.0730
Male	6–11	Low	0.0750	11	0.0759
Male	12–19	Low	0.0759	11	0.0759
Male	60–69	All	0.0845	10	0.0845
Male	20–29	Low	0.0915	9	0.0920
Female	60–69	Low	0.0920	9	0.0920
Male	30–39	Low	0.0920	9	0.0920
Male	70+	All	0.0965	8	0.0965
Female	40–49	Low	0.1099	7	0.1099
Female	50–59	Low	0.1342	6	0.1342
Female	3–5	All	0.1363	5	0.1370
Female	3–5	Low	0.1363	5	0.1370
Male	3–5	Low	0.1370	5	0.1370
Male	3–5	All	0.1370	5	0.1370
Male	70+	Low	0.1392	4	0.1392
Male	40–49	Low	0.1565	3	0.1565
Male	60–69	Low	0.1671	2	0.1671
Male	1–2	All	0.1991	1	0.2004
Male	1–2	Low	0.1991	1	0.2004
Male	50–59	Low	0.2002	1	0.2004
Female	1–2	Low	0.2004	1	0.2004
Female	1–2	Low	0.2004	1	0.2004

Table 7. Sampling messages by sex, income, and age, CSFII/DHKS 1994

Message (number)	Male		Female		Dwelling units assigned sampling message (proportion)
	All-income	Low-income	All-income	Low-income	
	------(Age)-----				
1	1-2	50-59	1-2		0.1663
2	1-2	50-69	1-2		0.0527
3	1-2	40-69	1-2		0.0863
4	1-2	40+	1-2		0.0111
5	1-5	40+	1-5		0.0138
6	1-5	40+	1-5	50-59	0.1215
7	1-5	40+	1-5	40-59	0.0667
8	1-5, 70+	40-69	1-5	40-59	0.0223
9	1-5, 70+	20-69	1-5	40-69	0.0374
10	1-5, 60+	20-59	1-5	40-69	0.0429
11	1-5, 60+	6-59	1-5	40-69	0.0147
12	1-5, 50+	6-49	1-5	6-11, 40-69	0.0105
13	1-5, 50+	6-49	1-5, 60-69	6-19, 40-59	0.0073
14	1-5, 50+	6-49	1-11, 50-69	12-19, 40-49	0.0216
15	1-11, 50+	12-49	1-11, 50+	12-19, 40-49	0.0249
16	1-11, 50+	12-49	1-11, 50+	12-49	0.0339
17	1-11, 50+	12-49	1-19, 50+	20-49	0.0093
18	1-19, 50+	20-49	1-19, 50+	20-49	0.0093
19	1-19, 40+	20-39	1-19, 50+	20-49	0.0218
20	1-19, 40+	20-39	1-19, 40+	20-39	0.0155
21	1-29, 40+	30-39	1-19, 40+	20-39	0.0062
22	1-29, 40+	30-39	1-29, 40+	30-39	0.0058
23	1+		1-29, 40+	30-39	0.0245
24	1+		1+		0.1735

Table 8. Number of SP's eligible for intake interviews; number completing day 1; and corresponding sample size targets by income, sex, and age, CSFII 1994–96

Sex and age (years)	Low-income households			All households		
	Eligible SP's in low- income households	SP's completing day-1 intake	CSFII 1994–96 low-income sample size target	Eligible SP's in all- income households	SP's completing day-1 intake	CSFII 1994–96 all-income sample size target
Males						
Under 1	69	61	NA	213	187	NA
1–2	252	245	207	803	725	719
3–5	257	238	207	850	734	719
6–11	225	215	207	867	751	719
12–19	233	218	207	881	734	719
20–29	262	229	207	1,017	779	793
30–39	243	201	207	1,157	890	850
40–49	237	195	207	1,138	861	850
50–59	231	204	207	1,186	888	850
60–69	223	202	207	1,092	846	850
70 and over	221	206	207	993	790	793
Females						
Under 1	79	73	NA	222	195	NA
1–2	246	237	207	794	707	719
3–5	250	238	207	834	735	719
6–11	220	214	207	841	734	719
12–19	234	216	207	876	732	719
20–29	256	236	207	960	726	739
30–39	221	207	207	963	809	793
40–49	247	226	207	1,142	903	850
50–59	202	188	207	1,071	864	850
60–69	224	209	207	1,001	790	793
70 and over	243	230	207	917	723	719
Total, excluding children < 1	4,727	4,354	4,140	19,383	15,721	15,482
Total	4,875	4,488	4,140	19,818	16,103	15,482

NOTE: Table excludes SP's who were selected for the survey but became ineligible before completing the day-1 intake. Classification by income, sex, and age reflects imputed values. Age is that at the time of screening.

Table 9. Number of SP's completing the day-1 intake interview and number selected for completing the DHKS interview, 1994–96

Income, sex, and age (years)	SP's completing day-1 intake*		SP's selected for DHKS**		SP's completing DHKS	
	(number)	(percent)	(number)	(percent)	(number)	(percent)
Low-income						
Males						
20–39	430	17	264	15	239	15
40–59	399	16	287	16	260	16
60 and over	408	16	297	16	259	16
Females						
20–39	443	17	297	16	270	16
40–59	414	16	309	17	293	18
60 and over	439	17	352	19	323	20
Total	2,533	100	1,806	100	1,644	100
All-income						
Males						
20–39	1,669	17	983	16	874	15
40–59	1,749	18	1,120	18	1,036	18
60 and over	1,636	17	1,080	17	987	17
Females						
20–39	1,535	16	933	15	847	15
40–59	1,767	18	1,119	18	1,047	18
60 and over	1,513	15	1,059	17	974	17
Total	9,869	100	6,294	100	5,765	100

* Includes all SP's completing the day-1 intake.

** Excludes SP's who became ineligible before completing the DHKS.