AUTOMATION OF WITHIN-HOUSEHOLD SAMPLING

Joseph Waksberg and Leyla Mohadjer, Westat, Inc.
Joseph Waksberg, 1650 Research Blvd., Rockville, MD 20850

1. Introduction

Although most statisticians like to carry out sample selection in a central office where the operations can be closely monitored, this usually cannot be done in field surveys requiring sampling within households. In most cases, within-household sampling involves simultaneous screening, sampling, and interviewing during the same visit to a household. The most frequent use of interviewer sampling is probably in those household surveys where one adult per household, randomly selected, is to be the respondent. Other studies involving within-household sampling include those with specified sampling rates within households, and those in which some types of household members are to be sampled at higher rates than other types, e.g., age groups sampled at different rates, married women selected at higher rates than unmarried.

The trend towards computerization of interviewing by means of CATI and CAPI has reduced the involvement of interviewers in sample selection operations. However, many surveys are still carried out as face-to-face interviews with paper and pencil, and methods of sample selection for such surveys are critical parts of survey methodology. This paper contains a technique that is very effective in almost all surveys requiring within-household sampling in such face-to-face interviews. We also describe two sample selection methods that can be used when certain, specific requirements are imposed on a survey.

A well-known procedure for within-household sampling is the use of sampling tables described by Kish (1949, 1965). A set of tables is prepared, each table specifying which person is to be selected for the sample in a one-person household, in a two-person household, etc. For each household size, the persons to be selected are rotated among tables so that over the set of tables for any household size all persons have the same probability of selection. A fairly limited number of different tables is sufficient to provide almost unbiased samples. The tables are arranged in systematic order and assigned to the sample households, which are also arranged in some kind of natural sequence. The interviewer then simply follows the instruction on the particular table assigned to each household to select the sample person. The interviewer does not have to carry out any sampling operation; he or she just chooses the person indicated by the table.

Although Kish's tables are set up for the selection of one person per household, he points out that it would not be difficult to generalize the system to two or more persons per household. However, the number of tables needed for a reasonable amount of randomization would have to be increased substantially. Similarly, the tables could be expanded to cover situations in which household members are given different probabilities of selection (e.g., males selected at twice the rate of females, or elderly persons oversampled). This approach would also require an increased number of tables that are even more complicated. It is doubtful that the increased complexity would be practical to manage.

2. Computer-Generated Message System

Westat statisticians have developed a procedure for sampling which can be used for a variety of non-CAPI field operations involving quite complex sampling requirements. The development of the procedure followed a suggestion by Dr. W. Pratt of NCHS, and its first use by Westat was for the second cycle of the National Survey of Family Growth (NSFG) carried out under his overall direction. Since then, Westat has used it for other household surveys in which within-household subsampling has to be carried out as part of a joint screening and interviewing operation. Its simplest form is when one person is to be chosen per household. It is also applicable in more complex situations, for example, when different probabilities are assigned to members of specified groups, e.g., females selected at a different rate than males, elderly persons oversampled, married persons at a different rate than unmarried and when there are different sampling rates for various types of households, such as Black, White or Hispanic. The procedure can be applied when either a specific number of sample persons per households is required or when prespecified sampling rates are to be followed.

The procedure (referred to as a computer-generated message system) can be viewed as an adaptation of Kish's tables, bringing them up to date through the use of computer technology. The basic idea is to determine a sample selection instruction for each household (or other sampling unit), which is expressed in a form that does not require knowledge of the composition of the household prior to the screening operation. These sampling instructions are randomized among households so that the desired probabilities of selection are adhered to. The instructions tell the interviewer who the sample person or persons are in each household. The interviewer thus does not have to carry out what one normally thinks of as a sampling operation.

The computer-generated message system can be described most clearly by examples of its application to satisfy various sampling requirements. The first is the relatively simple situation in which one person is to be selected, at random in each household. For this, a slight extension of the Kish tables is used. The computer generates a series of messages, each one looking like the following:

Table 1. Sample selection for one person per household

If the number of person in the household is:	Select the following listed person for the sample:
1	1st
2	1st
3	3rd
4	2nd
5	4th
:	:

The numbers in the second column are completely randomized among the messages, using either a random number generator or a form of systematic rotation of the numbers to assure that on each line all persons are shown the same number of times. One message is attached to the screening document or questionnaire for each household. In practice, systematic arrangement does not provide a much more exact distribution than unrestricted random ordering because the number of persons in a household will vary from household to household so that for any household size, the line numbers used in a particular survey are likely to be close to a random set. If an adult is to be selected in each household, the same tables can be used. The column headings are changed to read "adult" instead of "person." There can be as many lines on the table as can fit conveniently on a message, although 10 or 12 lines will cover all cases likely to be found in practice. Westat usually has the household identification for each sample household placed on a message to eliminate all ambiguity on how the messages are to be assigned. Additional messages are attached to blank household indicators when sample households can be added in the field operation.

Table 2 illustrates messages to be generated when one or two persons are to be selected, depending on the household size.

A slight variation of this table can be used when one person per household is to be selected, but a subset of households (e.g., minorities, those with members over 65 years of age, etc.) are to be over-sampled. Two methods are possible. One is to have two tables. The first table, illustrated as Table 3, indicates whether the household is eligible for the survey; the second, designates the sample persons.

Table 2. Sample selection for one or two persons per household

Number of eligible persons in household:	Choose line number(s):
1	1
2	1
3	3
4	1, 4
5	2, 3
6	5, 6
7	6, 7
8	4, 7
9	3, 9

If more than 9 eligible persons in household, contact supervisor for instructions.

Table 3. Household eligibility for the survey

If household is:	Household inclusion in survey:
Hispanic	Yes
Black, not Hispanic	Yes
Other	No

For example, if Hispanics are to be sampled at three times the rate of "other" households and Blacks at twice the rate, then: (a) the Hispanic line would always say "yes;" (b) in two-thirds of the messages for Blacks the line would say "yes;" in one-third of the messages the line for other would say "yes," in two-thirds it would say "no." This table would be followed by Table 1 which would be used to select persons in households included in the sample.

An alternate way of satisfying the same requirements is to consolidate Tables 1 and 3 into a single table as shown in Table 4. For Black and other households, "none" is randomized among the lines over the set of messages so that the expected proportion of cases with "none" is the same as when Table 2 is used.

Whether the combination of Tables 1 and 3 or Table 4 should be used appears to be largely a matter of taste. There may be an advantage to Tables 1 and 3 when early screening on race/ethnicity (or other criterion used for oversampling households) reduces the amount of within-household screening for households determined to be excluded from the sample.

As will be reported later, the decision to oversample Hispanics and Blacks by restricting the sample to one-third of the White households and two-thirds of the Black sample is a somewhat naive method of sample selection. A procedure with lower variances is described in Section 3.

As a third example, we consider a situation in which some household members are to be sampled at a higher rate than others. Assume that persons 65 years or older are to be taken twice as often as others. Table 5 illustrates how an appropriate sample message could be organized.

The messages are randomized so that on each line, persons 65 and over occur at a rate twice that of other persons.

Although the sampling message system is theoretically applicable to a wide variety of sampling situations, its main applicability is probably for within-household subsampling. Westat's experience with this system is restricted to selection of persons within surveys. The size of the messages may become unwieldy with larger sampling units such as institutions or schools.

As in any computer operation, quality control is vital in making the system work. A programming error which distorts the probabilities of selection would be disastrous and thus, careful checking of the output is advisable.

Table 4. Sample selection for one person per household when household types are sampled at different rates

If the number of persons in the household is:	Select the following listed person for the sample:		
	Hispanic	Black	Other
1	1st	None	None
2	1st	1st	1st
3	3rd	3rd	None
4	2nd	None	None
5	4th	2nd	3rd
:			
:			

Table 5. Sample selection for one person per household when persons 65 and over are sampled at a different rate than other ages

If the number of persons in the household is:		Select the following listed persons:	
65+	<65	65+	<65
0	1	–	1st
0	2	–	1st
0	3	–	3rd
:	:	:	:
:	:	:	:
:	:	:	:
1	0	1st	–
1	1	1st	–
1	2	–	2nd
1	3	–	3rd
:	:	:	:
:	:	:	:
:	:	:	:
2	0	2nd	–
2	1	1st	–

3. Special Sampling Requirements

Within household sampling methods exist that improve survey efficiency considerably. We describe two techniques that were developed by Joseph Waksberg and have been used by Westat in a number of surveys. Although the two techniques were initially developed to satisfy the specific requirements of two surveys, they are applicable to a wide variety of studies.

Reducing Variability in Sampling Rates When Classes of Households Are Sampled at Different Rates

In giving examples of sampling tables that can be used when certain types of households (e.g., Black, Hispanic) are to be oversampled simultaneously with the selection of one person per household, we commented that a more efficient procedure was possible. This method was first developed for use in the fourth cycle of the National Survey of Family Growth, NSFG (Rieger, et al, 1989).

To describe the procedure, we use the same example given earlier, Hispanic and Black households are to be oversampled by factors of 3 and 2, respectively, and one person is to be selected per sample household. Selection of one person per household results in highly variable rates of selection: persons in one-person households have a certain chance of selection, those in two-person households have only half that probability of selection, the probability of selection is one-third as great in three person households, etc. These variable probabilities increase the variances considerably for almost all estimates. An efficient procedure that can be used in such situations is described below.

When Hispanic households are oversampled by a factor of 3, we can proceed as follows. To achieve an oversampling of 3, we need to screen three times as many households as is necessary for the nonHispanic, nonBlack sample (referred to as "White and other"). The sampling tables shown earlier, subsampled one-third of the "White and other" households to achieve the appropriate sampling rates. An improved procedure which achieves the same sample sizes with lower variances is to consider all "White and other" households as being potentially eligible for screening. In one person "White and other" households, one third of the persons are retained; in two-person households, two thirds of the households are retained and one person selected; in households with more than two persons, all households are retained, and one person selected per household. The sampling rates for 1 person, 2 person, etc. households are then proportional to 1/3, 1/3, 1/3, 1/4, 1/5, etc. instead of 1/3, 1/6, 1/9, 1/12, 1/15, etc., when households are subsampled before selection of persons.

The procedure will produce a greater sample size than the initial plan because persons in large households are not subsampled as severely. It will probably be necessary to scale back all sampling rates, for example, by retaining 1/4 instead of 1/3 of the households with one-person, 1/2 of two-person households, etc. This will bring the sampling rates even closer together.

A similar method can be applied to Black households. Instead of retaining two thirds of the households, all are screened and two thirds of the one-person households are retained with all of the two or more person households retained. A scaling back of sampling rates similar to the one for "other" households will also be necessary. The reduction in variability of sampling rates for Black households will not be as great as for "other" households, but will be the maximum possible without increasing the number of screened households or taking more than one person per household.

Our illustration of the method assumed oversampling Blacks and Hispanics, but it obviously can be applied in oversampling other kinds of households for which screening is necessary, e.g., households containing babies, school children, smokers, etc. The greater the diversity of sampling rates, the greater are the possible reductions in variability of sampling rates. When sampling messages are used, Table 4 is applicable.

Minimizing the Number of Households with Sample Persons

Some sample surveys are not restricted to one person per household but specify subdomains of household members that are sampled at different rates. In such surveys, some households will have members in the sample, but other households will be completely excluded. Recent Westat surveys with these requirements include the National Health and Nutrition Examination Survey III (NHANES III) which has highly variable sampling rates for specified sex-age groups and a national study on smoking patterns which had different sampling rates for smokers and nonsmokers.

In some surveys with such requirements the efficiency of the operations will be improved if the sample persons are clustered within households to the maximum extent possible. In NHANES III, for example, earlier research indicated that as a consequence of the manner in which households are compensated for participating in the survey, response rates were likely to be higher for households with many sample persons than for households with few such persons. In other surveys, costs are reduced when there are fewer households in the sample even though the number of sample persons is the same. A sampling technique for maximizing the average number of sample persons per household (equivalent to minimizing the number of households with sample persons) was developed for use in NHANES III

(Ezzati, et al, forthcoming). A description of this technique that has general applicability follows.

Assume that a screening sample has been designated and persons are to be subsampled. The persons are classified into L subdomains, with subsampling rates $r_1 \ldots r_L$. The subdomains are ordered by subsampling rate so that $r_i \leq r_{i+1}$. We note that the screening is the minimum amount necessary to achieve the sample size for the rarest subdomain so that all households in the rarest subdomain are selected, e.g., $r_L = 1$.

Table 6 describes the subsampling procedure.

It can easily be seen that this procedure will produce the correct subsampling rates for all subdomains. Furthermore, it maximizes the number of sample persons per selected household.

The computer-generated message system can be used with this type of sampling. Messages are generated in the computer which instruct the interviewers on which subdomains in the household are to be included in the sample. The proportion of messages with each combination of subdomain is r_1, $r_2 - r_1$, $r_3 - r_2$, etc. These messages are randomized and attached to the questionnaire. An example of the messages used for NHANES III is shown in Section 4.

Table 6. Procedure for subsampling persons to minimize number of sample households

A. There are L person subdomains with sampling rates $r_1 \leq r_2 \leq \ldots \leq r_L$

B. The screened households (n) are divided into L random subsets: the proportions in the L sets

$$r_1,\ r_2 - r_1,\ r_3 - r_2,\ \ldots\ r_{i+1} - r_i\ \ldots\ r_L - r_{L-1}$$

C. Subsampling rule:

Subset	Size	Persons in households in sample
L	$(r_L - r_{L-1})n$	Persons in subdomain L
L-1	$(r_{L-1} - r_{L-2})n$	Persons in subdomains L and L-1
L-2	$(r_{L-2} - r_{L-3})n$	Persons in subdomains L, L-1, and L-2
:		
:		
2	$(r_2 - r_1)n$	Persons in subdomain L, L-1, . . ., 2
1	$r_1 n$	All persons

4. Examples of Computer-Generated Messages

National Adult Literacy Survey (NALS)

NALS is one of the surveys currently being carried out by Westat, under a subcontract to Education Testing Services, for the National Center for Education Statistics (NCES). NALS requires sampling one person per household if there are no more than three eligibles in the household. In households with four or more eligible members, two respondents are selected. Exhibit 1 shows an example of the form used to list eligible persons in a household. The message, shown in Table 2, provides instructions to the interviewer on which household member/members to select.

National Health and Nutrition Examination Survey III

A more complex set of messages with variable subsampling rates for different subdomains of the population is used for the National Health and Nutrition Examination Survey III (NHANES III) conducted for the National Center for Health Statistics (NCHS). The set of subdomains for which specified reliability was desired consisted of sex-age groups for three race/ethnicity populations in the U.S. The race/ethnicity, sex, and age groups comprised 52 separate domains, but for sampling purposes they were collapsed into 18 groups, with a unique sampling rate for each group.

A national, probability sample of households is currently being selected for screening with the screening rate designed to produce the desired number of sampled persons for the most difficult age-sex domain in the race/ethnic group. Persons in other age-sex domains are subsampled.

The subsampling used the procedure for maximizing the number of sample persons per household described previously. There are 18 sex-age-race/ethnicity collapsed domains corresponding to the domains for which separate sampling rates are used. Exhibit 1 shows the definition of the 18 domains. Although there are 18 domains, it is necessary to divide the sample into 20 random groups (denoted by

L in Table 5) to identify sample households in which no Black or White/other household members are selected. Within each random group, members of particular sex-age-race/ethnic subdomains are identified as potential sample persons, other members are excluded from the sample. This is accomplished in the following way.

Exhibit 1

Subdomain Groups Sampled at Different Rates

Code	Content of sex/age message	White	Black	Mexican-American
1	Males:	none	none	60+ yrs
	Females:	none	none	XXX
2	Males:	2-11 mos.	2-11 mos.	60+ yrs
	Females:	2-11 mos	2 mos-5 yrs	2 mos - 5 yrs
3	Males:	2-11 mos. 80+ yrs.	2 mos-5 yrs, 60+ yrs	2 mos-5 yrs, 60+ yrs
	Females:	2-11 mos	2 mos-5 yrs	2 mos-5 yrs, 60+ yrs
4	Males:	2-35 mos, 80+ yrs	2 mos-5 yrs, 60+ yrs	2 mos-11 yrs, 40+ yrs
	Females:	2-35 mos, 80+ yrs	2 mos-5 yrs, 60+ yrs	2 mos-19 yrs, 60+ yrs
5	Males:	2 mos-5 yrs, 70+ yrs	2 mos-11 yrs, 40+ yrs	all
	Females:	2 mos-5 yrs, 70+ yrs	2 mos-11 yrs, 60+ yrs	2 mos-19 yrs, 40+ yrs
6	Males:	2 mos-5 yrs, 50+ yrs	all	all
	Females:	2 mos-5 yrs, 70+ yrs	all	all
7	Males:	2 mos-19 yrs, 40+ yrs	XXX	XXX
	Females:	2 mos-19 yrs, 40+ yrs	XXX	XXX
8	Males:	all	XXX	XXX
	Females:	all	XXX	XXX

Each household is assigned two types of sampling messages, a sampling message related to the race/ethnicity, and the second message related to the age and sex of the members of the household. The household is first screened for race/ethnicity. The race/ethnicity message then informs the interviewers which household members in the three race/ethnic groups are to be in the sample. One of the following four messages is attached to each household screening questionnaire:

Content of Race/Ethnicity Message

1. List all persons in appropriate race/ethnicity tables
2. List only White/other and Mexican-American persons
3. List only Black and Mexican American persons
4. List only Mexican American persons

In households that satisfy the race/ethnicity message, the listing of individuals includes their sex and age. The sex-age message then is used to identify potential respondents. (In some cases, the message states that no household members are to be selected.) Each household gets one sex/age message for each of the race/ethnicity groups.

References

Ezzati, T., Massey, J.T, Waksberg, J., Chu, A., and Maurer, K. *Sample Design for the Third National Health and Nutrition Survey (NHANES III)*, National Center for Health Statistics, Series 2 Report, forthcoming.

Flyer, Paul, *Sample Design for Institutional Population Component*, report by Westat, Inc. to Agency for Health Care Policy and Research, April 1990.

Kish, Leslie, **Survey Sampling,** John Wiley and Sons, 1965.

Kish, Leslie, *A Procedure for Objective Respondent Selection within the Household*, JASA, Vol. 44, 1949.

Rieger, S., Sperry S., Sprankle R., and Judkins, D., *National Survey of Family Growth, Cycle IV*, report by Westat, Inc. to National Center for Health Statistics, Nov., 1989.