THE NATIONAL YOUTH SURVEY

MH 27552

LEAA 78-JN-AX-0003

Delbert S. Elliott, Principal Investigator (NIMH & LEAA)
Suzanne S. Ageton, Co-Principal Investigator (NIMH)
Brian A. Knowles, Assistant Project Director (LEAA)
Tim Brennan, Investigator (LEAA)
Rachelle J. Canter, Investigator (NIMH)
David Huizinga, Investigator (NIMH)

PROJECT REPORT NO. 2

DESCRIPTION OF THE NATIONAL YOUTH SAMPLE

by

D. HUIZINGA

August, 1978

Behavioral Research Institute

Boulder, CO

This description of sampling procedures consists of two major parts. The first provides a general description of the multistage sampling process. The second provides technical notes on the sampling methodology.[1]

GENERAL DESCRIPTION

AN OVERVIEW OF THE SAMPLING PROCESS

The sample design for the 1977 National Survey of Youth was based on a multistage area probability sample of households in the continental United States. The sampling units of each stage of selection are (1) primary sampling units (PSU's), which are large geographical areas; (2) secondary sampling units (SSU's), which are smaller geographical areas, within PSU's; (3) segments, which are portions of SSU's; and (4) households within segments. Extensive stratification was used in the first two stages of selection.

The sampling process consists of (1) defining and describing the PSU's, stratifying the PSU's, and selecting certain PSU's to be part of the sample; (2) within each selected PSU, defining and describing the SSU's, stratifying these SSU's, and selecting certain SSU's as part of the sample; (3) creating and selecting segments within each selected SSU; and (4) from lists of addresses of households within segments, selecting households. The probabilities of selection for each stage were established to provide a self-weighted sample (i.e., every household had the same probability of inclusion in the sample).

SELECTION OF PSU's

PSU Definition

A primary sampling unit was defined as an entire Standard

---

[1]The assistance of Dr. Martin Frankel in the design of these sampling procedures is most gratefully acknowledged.

Metropolitan Statistical Area (SMSA)[2] or a county or group of contiguous counties containing a minimum of 5000 households. To achieve this minimum, counties with less than 5000 households were combined with neighboring counties to form a PSU meeting this requirement.

In the county combining process the following process was employed. All counties containing fewer than 6000 households were candidates for combination and counties containing fewer than 5000 households were required to be combined with other counties to achieve the 5000 minimum. Counties in the 5000-6000 household range were combined only if they could be combined with other counties of less than 5000 households. The contiguity of combined counties was not an absolute requirement but was considered desirable. Fortunately, the contiguity condition was met in all combined county PSU's.

Sampling Procedures

To select PSU's for inclusion in the sample, a replicated zone sample (Demming, 1960; Kish, 1965) was employed. A systematic sample with a random start and an interval equal to the zone size was employed to obtain

---

[2] For the purpose of this frame, Standard Metropolitan Statistical Areas or SMSA's are those areas so designated by the Census Bureau for the 1970 Census, with the exception of Census Bureau defined SMSA's in the New England Census Division. In New England, the Census Bureau uses townships and other local boundaries to create SMSA's. To be consistent with the rest of the country, and because the updated (to 1975) number of households was only available by county, the New England SMSA's were redefined in terms of counties.

For the PSU sample frame, the definition of a New England SMSA was taken to be a county or group of contiguous counties containing at least one or a portion of a Census Bureau defined SMSA. Each single-county New England sample frame SMSA completely contained at least one Census Bureau SMSA and each county of a multiple county New England sample frame SMSA contained some part of a mutually shared Census Bureau SMSA. This definition resulted in the formation of 15 sample frame SMSA PSU's in the New England Division. These 15 PSU's account for the 26 Census Bureau SMSA's of the New England Division.

a probability proportional to size (PPS) selection of one PSU from each of 76 zones. The measure of PSU size was the estimated number of households contained in the PSU. The stratification of PSU's with this method is implicit and depends on an ordering of the PSU's in the sample frame.

Stratification

To take advantage of the systematic sampling procedure, the sampling frame of PSU's was stratified (ordered) on three major variables. These were (1) SMSA, Non-SMSA, (2) Census Division, and (3) size of PSU in terms of households. Also, in the South Atlantic division, percent black was used as a stratifying variable.

The actual ordering was as follows:

1. The frame was divided into an SMSA section and a Non-SMSA section. This provided an "urban-rural" split of the frame.
2. Within each of the SMSA and Non-SMSA sections, the frame was ordered by Census division. Each SMSA was considered as lying in only one Census division. In cases where an SMSA was divided between two or more Census divisions, the entire SMSA was assigned to the division in which the greatest percentage of the SMSA population resided. The Census geographic divisions were ordered in a serpentine fashion, as illustrated in Figure 1, thus insuring geographic stratification of the sample.
3. Within the South Atlantic divisions, the PSU's were further divided into those whose population was less than 20% black and those whose population was more than 20% black. The 20% criteria was established so that both black and non-black sections covered multiple zones. The other Census divisions were examined for a black/non-black split. Since the split criteria

Figure 1. PSU Frame Arrangement

Urban/Rural Split	Geographic Split	% Black Split	# Households Ordering
SMSA	New England		Ascending
SMSA	Mid Atlantic		Descending
SMSA	East North Central		Ascending
SMSA	West North Central		Descending
SMSA	Mountain		Ascending
SMSA	Pacific		Descending
SMSA	West South Central		Ascending
SMSA	East South Central		Descending
SMSA	South Atlantic	Greater than 20%	Ascending
SMSA	South Atlantic	Less than or equal to 20%	Descending
Non-SMSA	New England		Ascending
Non-SMSA	Mid Atlantic		Descending
Non-SMSA	East North Central		Ascending
Non-SMSA	West North Central		Descending
Non-SMSA	Mountain		Ascending
Non-SMSA	Pacific		Descending
Non-SMSA	West South Central		Ascending
Non-SMSA	East South Central		Descending
Non-SMSA	South Atlantic	Greater than 20%	Ascending
Non-SMSA	South Atlantic	Less than or equal to 20%	Descending

would have been less than 10% in order to cover multiple zones in these divisions, such splits were not made.

4. Finally, within each of those 20 sections, the PSU's were ordered in an ascending or descending sequence in a back to back manner on the basis of number of households. The actual arrangement is indicated in Figure 1.

Data Sources

The number of households per county was taken from the estimates for 1975 provided by Sales Management Magazine, July 1, 1975. The Census Bureau State and County codes, obtained from the Geographic Identification Code Scheme Booklet for the 1970 Census PHC(R)3, were used for county identification. The maps of Counties, SMSA's and Selected Places by State, published by the Census Bureau, were used for the geographical location of counties. Data for the proportion black was taken from the 1970 Census, published in the County and City Data Book - a Statistical Abstract Supplement (1972).

PSU Sample Frame and Sample Selection

The PSU sample frame described above contained 2009 PSU's; 231 SMSA PSU's and 1778 Non-SMSA PSU's. These PSU's accounted for a total of 3107 counties. The Census Bureau lists 3108 counties in the continental United States. This discrepancy results from the use of a combined Nanesmond County and independent Suffolk City, Virginia, by Sales Management Magazine, from which estimates of the number of households per county were obtained.

The total number of households contained in the frame was 70,940,900. The sample draw was based on the creation of 76 zones, so that to obtain an integral zone size, 1 blank household was added to the number of households

of each of the last 8 PSU's. This provides a zone size Z=933,433 households.

The sample draw employed a PPS systematic procedure with sample numbers $R+kZ$ $K=0,1,...75$, where R is a "random number", $1 \leq R \leq Z$.

Those PSU's that contain more households than the zone size entered the sample with certainty. Some of these PSU's "cover" several zones and could be "selected" more than once. For PSU's selected k times ($k>1$), k "replicates" were formed at the second stage.

SELECTION OF SSU'S

SSU Definition

Within each selected primary sampling unit (PSU), secondary sampling units were taken to be Block Groups (BG's) or Enumeration Districts (ED's), as defined by the Census Bureau for the 1970 Census, with the requirement that each BG or ED must contain at least 60 households. For purposes of the SSU frame, a household is defined as a dwelling that is habitable on a permanent basis and excludes seasonal and migrant housing units. Any BG or ED not meeting the above minimal requirement was combined with neighboring BG's or ED's to reach the 60 household minimum, and this combined BG/ED was taken as one secondary sampling unit. The number of households in a BG/ED was taken from the "first count data" of the 1970 Census.

In the combining process the following rules were applied: (1) BG's must be combined with BG's and ED's with ED's (this was done to insure that urban and rural areas could be separated, see below); (2) combined BG's or combined ED's must belong to the same census tract and to the same Minor Civil Division or Census Civil Division (this insures the combined areas are in the same general geographical area); and (3) the combined

areas should have sequential BG or ED census numeric identifiers. (This, in general, assures that the combined regions are contiguous, although there are a few cases where this is not the case.) The process of combining "undersized" BG's and ED's with neighboring BG's and ED's was automated. A description of the automated process and computer program which performed this process are contained in Appendix I.

Sampling Procedures

A SSU sample frame (one for each PSU) contains a sequence of SSU's, each representing one BG or ED or group of BG's or ED's meeting the minimum size requirement. A probability proportional to size, systematic sampling procedure with a random start was employed to select SSU's from the sample frame constructed for each PSU. The measure of size for each SSU was the number of households contained in the SSU according to the 1970 census. Six SSU's were selected from each previously selected PSU. For PSU's selected k times (k>1), 6k SSU's were selected. The k "PSU" selections being represented as follows: "PSU" selection j contains the $j+mk^{\underline{th}}$, m=0,1,2,...,5 selected SSU. Fractional zone sizes were employed in making the systematic PPS draw (see e.g., Kish, 1967, p.116).

Stratification

To take advantage of the systematic draw used in selecting SSU's, prior to selection, the SSU's from each PSU were ordered as described below. In this description, counties are the standardly defined political and administrative units. Minor Civil Divisions and Census Civil Divisions (MCD/CCD's) are subsets of counties, as are census tracts. Census tracts commonly are subsets of MCD/CCD's, but this is not always the case. BG's and ED's are subsets of census tracts. Full description of these various geographical units can be found in the Census Users Dictionary, 1970 Census

Users Guide, Part I, published by the Census Bureau. For the purpose of ordering the BG/ED's, if a BG or ED was split between more than one MCD/CCD, it was uniquely assigned to that MCD/CCD which contained the largest proportion of the BG/ED's households. The secondary unit sample frames were ordered as follows:

1. The SSU's (BG's and ED's) were arranged with all BG's first, followed by ED's.
2. Within each of these two divisions, the units were sorted into county groups and the county groups arranged by total county size (number of households), in decreasing order.
3. Within each county, the units were sorted into MCD/CCD groups and these MCD/CCD groups arranged by decreasing MCD/CCD size order.
4. Within each MCD/CCD group of units, the units were sorted by ascending census tract number and within census tracts by ascending block group number or enumeration district number. These identifiers are part of the Census Bureau Geographic Identification Scheme (see Census Users Guide, 1970, Part II). Rural ED's not assigned to a census tract were given a blank tract code.

For PSU's which contained a sufficiently large black population that an entire zone (interval of the systematic sampling procedure) or zones could be covered by BG's or ED's whose population was more than p% black, the SSU frame was first divided into two segments, those containing p% or less black and those with more than p% black. For rural areas p=20% and for highly urban areas p=50%. The above ordering process was then applied

independently to each of these two segments. The MCD/CCD and county household counts used for ordering were the MCD/CCD and county totals. These totals were also used in ordering the independent segments.

This ordering process provides an implicit stratification of the SSU frame based on (1) urban or rural characteristics; (2) general size in terms of the numbers of households in the local area containing a BG or ED; (3) geographical location; and (4) for some PSU's, ethnic distribution. The standard SSU frame arrangement is pictured in Figure 1.

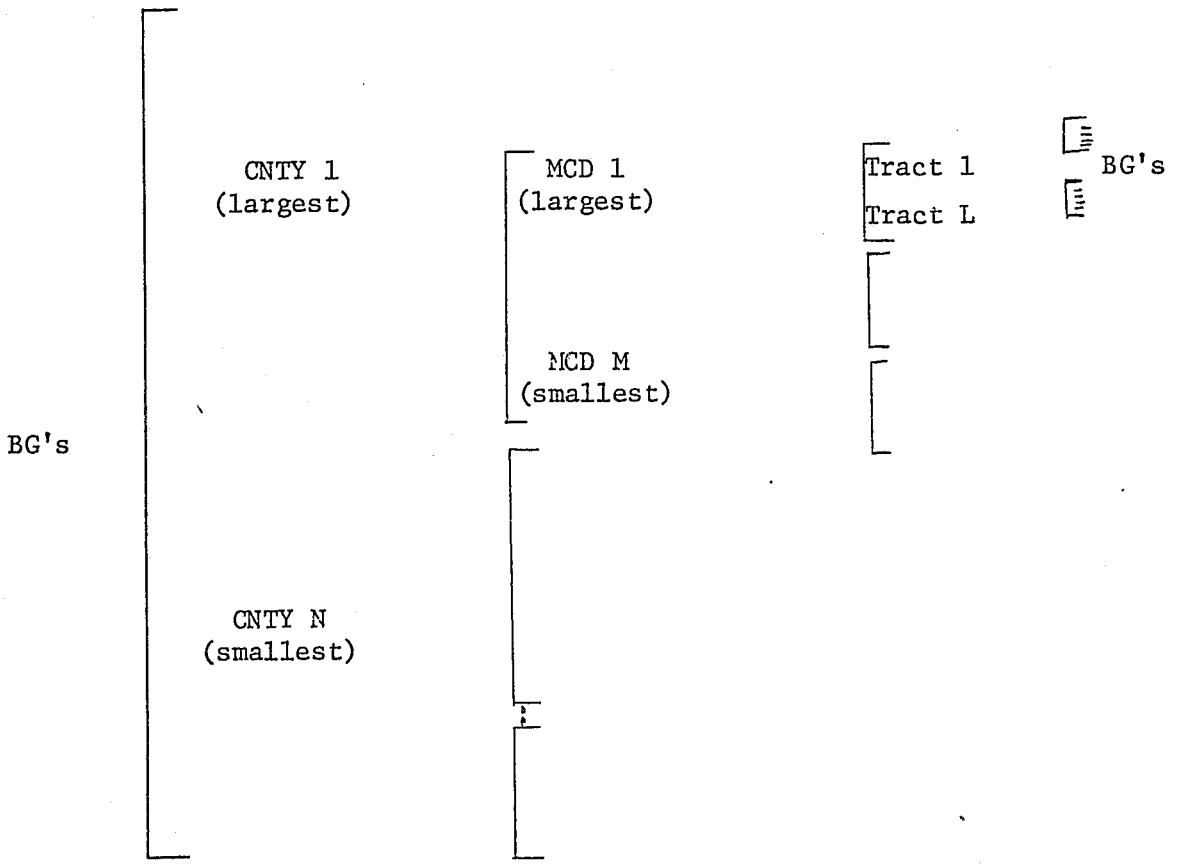
SELECTION OF SEGMENTS

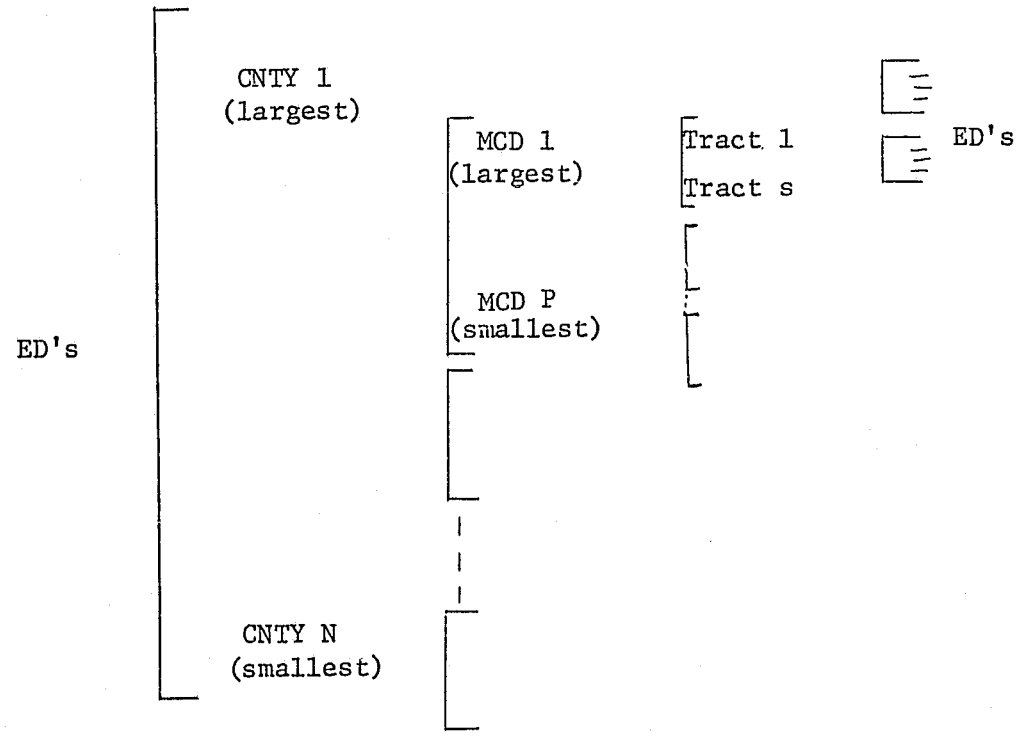
Segment Definition

Within BG's or ED's selected during the second stage, contiguous geographical areas of approximately 100 households were created. Two processes were used in creating the segments. For urban areas for which there are published block statistics, the segmentation proceeded "in-house," since all needed information (maps, households counts) were available. For areas for which there was no published data, enumerators performed "field counts" of the areas and these field counts were used for segmentation purposes. Although segments generally contained 100 households, a minimum of 60 households was permitted. SSU's with 60-100 households were not segmented.

In some instances, a segment selected from a BG or ED contained several hundred households. This resulted from either population growth in the segment since the 1970 census (1970 census data was used for "in-house" segmentation) or because one block, the smallest segment for which published information was available, contained several hundred households. In this case, a fourth stage of sampling was employed. These large segments were field counted and subsegments of approximately 100 households were created. One of these subsegments was then selected.

FIGURE 1


CNTY 1
(largest)
MCD 1
(largest)
Tract 1
Tract L
BG's
MCD M
(smallest)
BG's
CNTY N
(smallest)


CNTY 1
(largest)
MCD 1
(largest)
Tract 1
Tract s
ED's
MCD P
(smallest)
ED's
CNTY N
(smallest)

Sampling Procedure

One segment was selected from each SSU with probability proportional to size (PPS). Selection of a subsegment, when required, was also performed using PPS sampling.

SELECTION OF HOUSEHOLDS

Segments selected during the third or fourth stage of sampling were completely enumerated. From the resulting lists of households, a systematic sample of households was selected. The sample rates within segments were determined so that the entire sample of households was self-weighting. Determination of the sample rates is discussed in the technical section of this document.

A DESCRIPTION OF COMPLETION RATES AND A CHECK ON THE REPRESENTATIVENESS OF THE SAMPLE

The above sampling procedures resulted in the listing of 67,266 households in 456 segments. From this listing, approximately 8000 households were selected for inclusion in the sample. All 11 through 17 year old youth living in the selected households were the eligible respondents for the study[3]. An attempt was made to interview each youth and one of the youth's parents.

Of the selected households, 379 were vacant and in 59 of the households no occupant was ever found at home. Among households in which an

---

[3]The approximately 8000 household sample size was determined to provide a sample of approximately 2100 eleven through seventeen year old youth. This number of households was based on assumptions of a 7% vacant household rate, a 75% completion rate of occupied households, and an average of 0.38 11-17 year old youth per household.

occupant was located, 6117 households did not contain eligible youth and in 34 households, respondents refused to participate in the study and would not provide information about household members. In 353 of the households containing youth in the appropriate age range, parents refused to allow their youth to participate in the study. In most cases, these parents did indicate the number of eligible youth living at home and it is estimated that these households contained 610 eligible youth. The remaining 1056 households contained 1765 eligible youth. Of these, 19 refused to participate in the study and 20 were considered ineligible for the study for reasons such as mental retardation. Interview schedules were completed for the remaining 1726 youth.

Parents of youth respondents were also interviewed. Of 1056 potential parent respondents, 17 refused to participate in the study, although allowing their youth to participate.

The completion rates described above are given in tabular form in Table 1.

TABLE 1

COMPLETION RATES RESULTING
FROM THE NATIONAL SURVEY OF YOUTH

Households	
Number of Households in the Sample	7998
Households Not Interviewed	
Vacancies	379
Not at Homes	59
Refusal by Parents at the door Eligible youth live in household.	353
Refusal by adult respondent Whether eligible youth live in household is unknown.	34
Households Interviewed	
Households with no eligible youth	6117
Households with eligible youth	1056
Eligible Youth Respondents	
Estimated number of youth not interviewed because of parent refusal.	610
Number of youth refusing to participate in study.	19
Number of youth considered inappropriate for inclusion in the study.	20
Number of youth that completed interviews.	1726
Total number of eligible youth	2375
Completion rate among eligible youth respondents.	73%

The age, sex and ethnicity characteristics of the youth sample are presented in Table 2. In that table, they are contrasted with recent estimates provided by the Census Bureau for the total 11-17 year old youth

population. The age, sex, and ethnicity of eligible youth not interviewed (for those youth for whom such information is known) is presented in Table 3. As indicated in the table, the loss rate from any particular group appears, in general, to be proportional to that group's representation in the population. Thus, on the basis of demographic characteristics, the sample appears to be representative of the total 11-17 year old youth population.

TABLE 2

DEMOGRAPHIC CHARACTERISTICS OF THE NATIONAL YOUTH SURVEY SAMPLE AND OF THE NTOTAL 11-17 YEAR OLD POPULATION

	SAMPLE	CENSUS BUREAU* POPULATION ESTIMATES
ETHNICITY		
Anglo/Chicano	83%	84%
Black	15%	14%
Other	02%	02%
SEX		
Male	53%	51%
Female	47%	49%
AGE		
11	13%	13%
12	14%	14%
13	16%	14%
14	14%	15%
15	16%	15%
16	14%	14%
17	13%	15%

* Source: Population Estimates and Projections, Series P-25, No. 643. Bureau of the Census, 1977

TABLE 3

DEMOGRAPHIC CHARACTERISTICS OF ELIGIBLE YOUTH NOT INTERVIEWED

	AGE							SEX		ETHNICITY			
	11	12	13	14	15	16	17	Male	Female	Anglo	Black	Chicano	Other
PARENT REFUSAL to allow youth to participate	67	82	71	65	61	69	70	188	186	271	50	19	16
YOUTH REFUSAL	1	1	0	4	0	6	2	10	5	11	0	0	0
YOUTH INAPPROPRIATE for interviewing	3	6	5	1	4	4	3	13	9	4	2	1	0
TOTAL	71	89	76	70	65	79	75	211	200	286	52	20	16

TECHNICAL NOTES

As described above, primary sampling units (PSU's), secondary sampling units (SSU's) and segments were selected with probability proportional to size. With the exception of logistical concerns, these selection procedures are straightforward and will not be discussed in further detail. In the following, the determination of sample rates within segments, estimators for proportion and frequency estimates, and variances for these estimates are described.

DETERMINATION OF SAMPLE RATES WITHIN SEGMENTS

Based on the sampling procedures described above, it is desired to find an appropriate number of households to select from each final segment, in order to insure an equal probability of selection for all households. In the following, the measure of size (MOS) of a unit is the estimated number of households contained in that unit.

Let PSU_i have MOS M_i, $i=1,2,\ldots,76$;

$BG_{ij} \subseteq PSU_i$ have MOS N_{ij}, $j=1,2,\ldots,a_i$;

$Segment_{ijk} \subseteq BG_{ij}$ have MOS Q_{ijk}, $k=1,2,\ldots,b_j$;

and Subsegment ijkl $\subseteq$ segment ijk have MOS S_{ijkl}, $l=1,2,\ldots,d_k$.

In cases where the final stage units are block groups, subsegment ijkl = segment ijk = BG_{ij}, and in cases where the final stage units are segments,

subsegment ijkl = segment ijl.

$$\text{Let } P_{ijkl} = \frac{M_i}{\left(\frac{\Sigma M_i}{76}\right)} \quad \frac{N_{ij}}{\left(\frac{\Sigma N_{ij}}{6}\right)} \quad \frac{Q_{ijk}}{\Sigma Q_{ijk}} \quad \frac{S_{ijkl}}{\Sigma S_{ijkl}}$$

$$= \frac{76M_i}{\Sigma M_i} \quad \frac{6N_{ij}}{\Sigma N_{ij}} \quad \frac{Q_{ijk}}{\Sigma Q_{ijk}} \quad \frac{S_{ijkl}}{\Sigma S_{ijkl}} .$$

Since one PSU was selected from each of 76 zones, one SSU selected from each of 6 equal sized intervals, and one segment and one subsegment selected from the segment and subsegment frames, all with the probability proportional to size, P_{ijkl} is the probability of selecting subsegment ijkl.

Let Y_{ijkl} be the sampling rate within subsegment ijkl. It is desired that $P_{ijkl}\ Y_{ijkl} = C$, a constant, for all i,j,k, and l. Let

$$Y_{ijkl} = \frac{\Sigma N_{ij}}{M_i} \quad \frac{\Sigma Q_{ijk}}{N_{ij}} \quad \frac{\Sigma S_{ijkl}}{Q_{ijk}} \quad \frac{K}{S_{ijkl}} ,$$

for some fixed constant K. Then the probability of selecting any given household becomes

$$P(\text{household}) = P_{ijkl}\ Y_{ijkl} = \frac{76 \cdot 6 \cdot K}{\Sigma M_i} = C$$

and the sampling rate Y_{ijkl} is

$$Y_{ijkl} = \frac{C}{P_{ijkl}} .$$

Now let E_{ijkl} be the true or enumerated size of subsegment ijkl and let T be the total desired sample size. Then it is required that

$$\Sigma\, Y_{ijkl}\ E_{ijkl} = T$$

$$\text{or} \qquad \Sigma\, \frac{C}{P_{ijkl}}\, E_{ijkl} = T$$

$$\text{so that} \qquad C = \frac{T}{\Sigma\, \frac{E_{ijkl}}{P_{ijkl}}} .$$

Thus, to insure equal probability of selection for households and to obtain the desired total sample size T, the sampling rate for subsegment ijkl is $Y_{ijkl} = \frac{C}{P_{ijkl}}$, and the corresponding sample size H_{ijkl} is given by $H_{ijkl} = Y_{ijkl}\, E_{ijkl} = \frac{C}{P_{ijkl}}\ E_{ijkl}$.

In general, H_{ijkl} will not be an integer. In the following a procedure which determines an integral value for H_{ijkl} and maintains the same overall sampling fraction is described. Since the procedure is the same for all subsegments, for convenience the subscripts are ommitted, i.e. $E = E_{ijkl}$ and $H = H_{ijkl}$. Let $H = N + f$, where N is an integer and f a decimal fraction. If $f = 0$, the H is integral and the number of households to select is N. Suppose $0<f<1$. It is desired to find p such that $P\left(\frac{N}{E}\right) + (1-p)\left(\frac{N+1}{E}\right) = \frac{N+f}{E}$. Solving for p, $p = 1-f$ or $1-p = f$. Thus if N is chosen as the number of households to select with probability 1-f and if N+1 is chosen as the number of households to select with probability f, then

$$\begin{aligned} P(\text{household}) &= P(N \text{ households are to be selected}) \times \\ &\qquad P(\text{household} \mid N \text{ to be selected}) \\ &\quad + P(N+1 \text{ households are to be selected}) \times \\ &\qquad P(\text{household} \mid N+1 \text{ to be selected}) \\ &= (1-f)\frac{1}{\left(\frac{E}{N}\right)} + f\frac{1}{\left(\frac{E}{N+1}\right)} \\ &= \frac{N+f}{E} . \end{aligned}$$

It follows that if $H_{ijkl} = N+f$, N an integer and $0<f<1$, and a random number $0<R<1$ is selected, and if $R \geq f$ let $H_{ijkl} = N$ and if $R<f$ let $H_{ijkl} = N+1$, then an integral sample size is determined which maintains the same overall sampling fraction.

ESTIMATES, VARIANCES, AND CONFIDENCE LIMITS

The sampling procedures described above result in an equal probability of selection sample of households and since all 11 through 17 year old youth in a selected household were interviewed, an equal probability of selection of 11-17 year old youth (living in households). It is desired to estimate R, the proportion of 11-17 year old youth in the population that have performed a particular behavior or that have a particular characteristic. An estimate of the total number of times that 11-17 year old youth perform certain behaviors is also desired.

It is convenient to consider the overall sampling plan as resulting in 38 strata or zones, each comprised of two adjacent "half strata", thus accounting for the original 76 zones.

Let X be the total number of 11-17 year old youth in the population and Y be the number of such youth with a given characteristic. Let $X_h = X_{h_1} + X_{h_2}$ and $Y_h = Y_{h_1} + Y_{h_2}$ be the sample stratum totals with h_i denoting the half-stratum totals. Let N_h be the size of stratum h and n_h the sample size from stratum h. An estimate r of R is the combined ratio estimate

$$r = \frac{\Sigma N_h \bar{y}_h}{\Sigma N_h \bar{x}_h} = \frac{\Sigma (N_h/n_h) y_h}{\Sigma (N_h/n_h) x_h} = \frac{\Sigma y_h}{\Sigma x_h} .$$

The final term results from using a uniform sampling fraction in all strata. Although r is a biased estimate of R, the bias is less than the coefficient of variation of x (see e.g. Kish, 1967, p. 208). As a rule of thumb, it is often required that the coefficient of variation of x be less than 0.1 in order to control the degree of bias and to insure the adequacy of the estimated variance of r, described below.

(1/f)

An estimate of the variance of r, where variance refers to the mean square error about R, is obtained by assuming that the sample design and actual sample are such that $(1/f)x \doteq X$. (Comments on the effect of this approximation on the variance of R can be found in Kish, 1967, pp.207-8). Then letting V and Cov denote variance and covariance respectively, and $(1/f)x = \hat{X}$, $(1/f)y = \hat{Y}$

$$V(r) = E(r-R)^2 = E[(y/x) - R]^2 = E[(\hat{Y}/\hat{X}) - R]^2$$

$$\doteq \frac{1}{\hat{X}^2} E(\hat{Y}-R\hat{X})^2$$

$$= \frac{1}{\hat{X}^2} E[(\hat{Y}-Y) - R(\hat{X}-X)]^2$$

$$= \frac{1}{x^2} [V(y) + R^2V(x) - 2R\,Cov(x,y)] .$$

To obtain an estimate of V(y), the method of collapsed strata is used. Since identical designs are used in all half-strata, $V(y) = V(\Sigma y_h) = \Sigma V(y_{h_1} + y_{h_2})$. Ignoring the finite population correction, an estimate of $V(y_{h_1} + Y_{h_2})$ is $(y_{h_1} - y_{h_2})^2$, which provides a slight overestimate of the variance, the degree of overestimation depending on the term $[E(y_{h_1}) - E(y_{h_2})]^2$. Employing similar estimates for V(x) and Cov(x,y) and using the sample r for R, an estimate v(r) of V(r) is

$$v(r) = \frac{1}{x^2} [\Sigma(y_{h_1} - y_{h_2})^2 + r^2 \Sigma(x_{h_1} - x_{h_2}) - 2r\Sigma(x_{h_1} - x_{h_2})(y_{h_1} - y_{h_2})]$$

$$= \frac{1}{x^2} \Sigma[(y_{h_1} - y_{h_2}) - r(x_{h_1} - x_{h_2})]^2 .$$

The above is sufficient to estimate the proportion of 11-17 year old youth with a given characteristic and to estimate the sampling variance of this proportion. To estimate the total number of such youth, let X be the total number of 11-17 year old youth in the population and let r be the estimated proportion of youth. Then $N = rX$ provides an estimate of the total number of youth with the given characteristic and $v(N) = X^2 v(r)$ provides an estimate of the variance of N where v(r) is determined above.

To obtain an estimate of the total number of times a particular behavior is performed by 11-17 year old youth, the above ratio estimator can be employed. Taking y to be the total number of behaviors performed by youth in the sample, x to be the number of youth in the sample, and using similar definitions of y_h, y_{h_i}, x_h, x_{h_i}, r becomes an estimate of the average number of times the behavior is performed by a youth. Then, letting X be the total number of youth, $N_B = rX$ provides an estimate of the total number of times the behavior is performed and $v(N_B) = X^2 v(r)$, where v(r) is determined as above, provides an estimate of the variance of N_B.

Confidence limits for any of the estimates are obtained by assuming that the ratio estimates are approximately normally distributed. Since there are 38 full strata employed in the sample design, there are 38 degrees of freedom associated with the variance estimates. Thus, for an α level confidence interval, let $t_{\alpha/2}$ be the value of Students t distribution such that $P(t \geq t_{\alpha/2}) = \alpha/2$. Then

$$P(r - t_{\alpha/2} \sqrt{v(r)} \leq R \leq r + t_{\alpha/2} \sqrt{v(r)}) = 1 - \alpha$$

Similar expressions hold for N and N_B, employing v(N) and $v(N_B)$, respectively.

ADEQUACY OF RATIO ESTIMATES AND DESIGN EFFECTS

The adequacy of the ratio estimates and the variance of these estimates is dependent on the coefficient of variation (cv) of the x variable. Based

on a large number of calculated proportion estimates, in general the cv(x) for each estimate is less than 0.08 with a range of .05 to .08. The estimated design effect for these estimates (ratio of the obtained variance to the variance that would be obtained with a simple random sample) is generally in the range 1.00 to 2.50 with a mean design effect of 1.35.

APPENDIX A

SPECIAL TECHNIQUES RELATED TO SECOND STAGE SAMPLING

Within each PSU, secondary sampling units (SSU's) were taken to be Block Groups (BG's) or Enumeration Districts (ED's) with the requirement that each BG or ED must contain at least 60 households according to the 1970 census(all households were counted except seasonal and migrant). Any BG or ED not meeting this requirement was combined with neighboring BG's or ED's, respectively, to reach the 60 household minimum and this combined unit was taken as one SSU.

Source of Data

The 1970 census BG/ED household counts, as well as other information about BG's/ED's , is available from data known as "first count data" provided by the Census Bureau. These data are available on magnetic tape from either the Census Bureau or from private Summary Tape Processing Centers recognized by the Census Bureau (see Part I of the 1970 Census Users Guide).

To reduce the total cost of required data, only that information about each BG/ED needed for sampling purposes (and not the complete "first count data") was obtained, and these data were obtained only for those counties contained in the PSU's included in the first stage sample.

Creating the Sample Frame

In the following, a brief description of the automated process used in creating the second stage sample frame is described.

The "first count data" obtained include: (1) geographical identification consisting of state, county, MCD/CCD, census tract and BG or ED identifiers, and (2) population and housing data. Because the Census Bureau provides

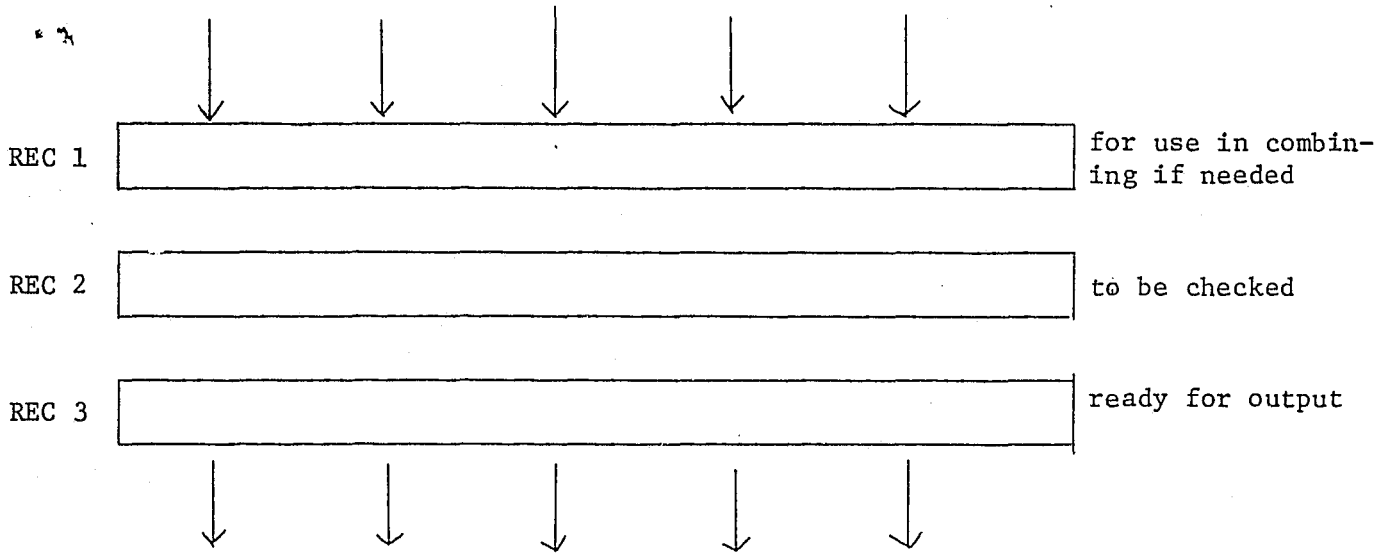
summaries for geographical areas which cut across BG and ED boundaries (e.g., political subdivisions), each BG or ED may be represented by several records of data. For sampling purposes, where BG/ED's are the elements of the frame, this requires all data representing one BG or ED be integrated into one record. Because all records for one BG or ED are contiguous in the Census Bureau ordering, this requires sequentially combining all records that have the same BG or ED identifiers, census tract, county and state codes. This combining simply involves adding corresponding data values from the records representing one BG or ED.

Because the MCD or CCD number is to be used in a later stratification process and because MCD/CCD boundaries may "split" BG's and ED's, each BG or ED was given a unique MCD/CCD identifier by assigning the BG or ED to that MCD/CCD which contained the largest proportion of the BG/ED population.

Following this initial combining process, each BG or ED is represented by one record. These records were then sorted into groups, one group per PSU, with the records maintained in Census Bureau order.

A secondary sampling unit was required to contain a minimum of 60 households. To insure that this requirement was met, the data for each PSU were examined to locate any BG's or ED's with less than this minimum size. Such BG/ED's were combined with other BG/ED's by the following automated process. Except as noted, all combined BG/ED's are from the same tract, MCD/CCD, and county.

The actual combining process consists of looking at three records as pictured below. The third record is ready for output, having size greater than 60. The middle record is the record being checked, and the first is present for use in the combining process if needed.



If record 2 represents a BG/ED with size $\geq$ 60 record 3 is output, record 2 moves to record 3 position and a new record is brought in. If record 2 represents a BG/ED with less than 60 households the following rules are applied. Except as noted below, because of the census ordering either record 1 or record 3 (or both) are from the same county and tract as the middle to be combined-record.

(1) If only one of records 1 and 3 have the same county, MCD/CCD, and tract as record 2, it is combined with that record.

(2) If both records 1 and 3 have the same county, MCD/CCD, and tract as record 2, it is combined with the record having the smallest MOS.

(3) If records 1 and/or 3 have the same county and tract codes, but neither records 1 or 3 have the MCD/CCD code as record 2, then above rules are applied (with the exception of MCD/CCD requirement) and the combined units DU's households.

Following the combining stage, the appropriate records are moved and a new record 2 brought in for checking; e.g., if records 1 and 2 are combined and the combined record contains 60 or more households, record 3 is output, the combined record is put in record 3 and two new records are brought in.

A "hitch" in the method occurs when the remainder of a census tract, not

already output, contains fewer than 60 households. In this case units must be combined across tract boundaries. For later sorting, the combined unit is assigned to the tract containing the largest portion of the combined units' households.

In order to use the above automated combining process, it is necessary to insure that the SSU's so created can be correctly identified on maps and eventually by personnel in the field. Thus, for any combined unit it is necessary that (1) a listing of all original BG's and ED's making up 2 combined units be maintained, and (2) within the listing of combined BG's/ED's any changes in MCD/CCD number or tract number be clearly indicated. For this study, each record representing a combined unit was "flagged" and special flags for MCD/CCD or tract changes were also present in the record. For each combined unit, a special output was created that listed all the combined units.

A second stage sample frame (one for each PSU) thus consists of a sequence of records, each record representing one SSU and representing one BG or ED or group of BG's or ED's meeting the minimum size requirement. These records are then sorted as described in the selection of SSU's section of this document. Thus, the frame is arranged with:

(1) All BG's first and all ED's second.

(2) Within these two groups, records are arranged by county, and the counties are placed in decreasing size order.

(3) Within counties, records are sorted by MCD/CCD and the MCD/CCD's placed in decreasing size order.

(4) Within MCD/CCD groups, the records or SSU's are sorted by tract, with tracts placed in sequential order, and within tracts by

sequential BG or ED order. (N.B. Only a few of the ED's actually are in tracted areas, so the sort by tract for ED's commonly has little effect. Its use, however, allows a consistent, automated process.)

If a PSU is more than p% in black population (p=50% for heavily urbanized areas, p=20% otherwise), the SSU frame was first divided into black and non-black sections and the above process applied independently to each section. To determine what value of p to use in creating the sections for each PSU, the zone size or sample interval of the systematic sample was calculated and p selected so that the black section of the frame covered one or more zones.

Following the ordering process, proportionate to size, systematic sampling of the SSU's is a straightforward procedure. Selecting and listing the chosen SSU's is easily automated, although care must be taken to list all the special indicators contained in the records describing the SSU's.