Survey of Income and Program Participation

Working Paper Series

No. 8406

Papers Presented in the
Survey of Income and Program Participation
Session III
at the Annual Meeting of the
American Statistical Association
in Philadelphia, PA
August 13-16, 1984

(No. 8406)

U.S. Department of Commerce
Malcolm Baldrige, Secretary
Clarence J. Brown, Deputy Secretary
Sidney Jones, Under Secretary for
Economic Affairs

BUREAU OF THE CENSUS
John G. Keane,
Director

U.S. BUREAU OF THE CENSUS
John G. Keane, Director
C.L. Kincannon, Deputy Director
William P. Butz, Associate Director
for Demographic Fields

POPULATION DIVISION
Roger A. Herriot, Chief

Acknowledgements

This publication is composed of papers prepared by many different authors for presentation at the American Statistical Association on August 13-16, 1984. We would like to thank these authors for their cooperation in making the papers available for publication. Clerical and editorial assistance was provided by Hazel Beaton, Mary Kisner, and Delma Frankel.

Suggested Citation

"Papers Presented at the "Survey of Income and Program Participation session III at the annual meeting of the American Statistical Association in Philadelphia, PA, August 13-16, 1984," SIPP Working Paper Series No. 8406. U.S. Bureau of the Census, Washington, D.C. 1984.

Preface

This report is comprised of five papers featured in the "Survey of Income and Program Participation" session III, one of two in the Survey Research Methods Section of the annual meeting of the American Statistical Association.

This session covered a range of topics, both methodological and substantive, about longitudinal surveys and the Survey of Income and Program Participation (SIPP).

SIPP is a new Census Bureau survey collecting data that will help measure income distribution and poverty throughout the country more accurately. These data will be used to study Federal and state aid programs (such as food stamps, welfare, Medicaid, and subsidized housing), to estimate future program costs and coverage, and to assess the effects of proposed changes in program eligibility rules or benefit levels.

Households in the survey will be interviewed at 4-month intervals over a period of 2 1/2 years. The reference period will be the 4 months preceding the interview. In all, about 20,000 households will be interviewed, approximately 5,000 each month. Field operations will be handled through our 12 regional offices.

Recurring questions will deal with employment, types of income, and noncash benefits. Periodic questions will be added dealing with school enrollment, marital history, migration, disability, and other topics. Special supplemental questions will also be added to the SIPP questionnaire.

These papers discuss SIPP and its predecessor, the Income Survey Development Program (ISDP), an experimental program designed to test procedures used in conducting SIPP.

Session: Survey of Income and Program Participation III

Session Chair: Daniel Horvitz
Research Triangle Institute

Papers:

"Obtaining a Cross-Sectional Estimate from a Longitudinal Survey: Experiences of the ISDP."
Written by H. Huang, Bureau of the Census.

(Examines alternative cross-sectional weighting schemes.)

"Weighting of Persons for SIPP Longitudinal Tabulations."
Written by L. Ernst, D. Hubble, D. Judkins, D. B. McMillen, and R. Singh, Bureau of the Census.

(Discusses changes in the sample during the survey and various weighting methods.)

"Longitudinal Family and Household Estimation in SIPP."
Written by L. Ernst, D. Hubble, and D. Judkins, Bureau of the Census.

(Compares weighting approaches for use with longitudinal household and family concepts.)

"Early Indications of Item Nonresponse in SIPP."
Written by J. Coder and A. Feldman, Bureau of the Census.

(Studies nonresponse rates in the first SIPP interviews.)

Obtaining Cross-sectional Estimates from a Longitudinal Survey: Experiences of the Income Survey Development Program

I. INTRODUCTION

In 1975 the Secretary of the Department of Health, Education and Welfare (The Department of Health and Human Services (HHS) predecessor agency) authorized a program, the Income Survey Development Program (ISDP), to resolve technical and operational issues for a major new survey -- the Survey of Income and Program Participation (SIPP). Much of the work of the ISDP centered around four experimental field tests that were conducted in collaboration with the Bureau of the Census to examine different concepts, procedures, questionnaires, recall periods, etc. Two of the tests were restricted to a small number of geographic sites; the other two were nationwide. Of the two nationwide tests, the more important data collection was called the 1979 Research Panel. This panel consisted of nationally representative samples which provided a vehicle for feasibility tests and controlled experiments of alternative design features. Information concerning the ISDP may be found in Ycas and Lininger (1981), David (1983), and the survey documentation now available through the National Technical Information Service (1983).

The 1979 Research Panel was a multiple frame sample consisting of a general population (area) sample of 9300 initially designated addresses drawn from the 1976 Survey of Income and Education (SIE) and some Census

Bureau's current survey reserve measures and new construction update, and two list frame samples of (a) eligible applicants for the Basic Educational Opportunity Grant (BEOG) Program and (b) blind and disabled Supplemental Security Income (SSI) recipients.

The 1979 Research Panel was a longitudinal survey consisting of six waves of interviewing; one third of the panel was interviewed each month with subsequent interview for a given unit occurring every three months. A sample of addresses was chosen and persons living in the sample units (addresses) during the first wave of interviews were defined as original sample persons. For interviews subsequent to the first, the sample of addresses became a sample of persons; accordingly, original sample people were followed to their new addresses in subsequent interviews (with reasonable geographic constraints -- within 50 miles of any ISDP Primary Sampling Unit). Personal interviews were conducted in Wave 1 with all adults (persons sixteen years old and over) at the sampled address. These become the original sample persons. During Waves 2-6 all persons currently residing with an original sample person were interviewed. This means, for example, that if an original sample person moved to a new address with four other adults, then questionnaires were administered to everyone at the original sample person's new address. If any original sample person remained at the first wave address, anyone who moved into that address with the original sample person were also interviewed. Thus, interviews were conducted with all adults at an address as long as at least one of the adults present was an original sample person. Because of the ISDP rules, persons can be lost from

sample because they move beyond the survey's boundaries; in addition, people were added to the sample because they became part of the housing unit in which the original sample person resides.

Obviously, the universe changes continuously through the life of the survey. A great deal of interest exists, however, in developing cross-sectional estimates at the time of each interview wave. In the absence of drawing a new sample at each interview, any cross-sectional estimates developed for Waves 2-6 are subject to a population coverage bias. This paper will focus only on the covered population and present some unbiased base weights for cross-sectional estimators for the non-institutionalized U.S. population represented by the longitudinal sample (the population coverage bias will remain, however). Since the methodology for treating both area sample and list frame samples was needed for ISDP 1979 Research Panel, both will be described below. The estimation methods described here are directly applicable to the Survey of Income and Program Participation (SIPP), an overall description of which is found in Nelson, McMillen, and Kasprzyk (1984) and Herriot and Kasprzyk (1984).

II. THE POPULATION FOR CROSS-SECTIONAL ESTIMATES

We begin by defining the general population for which estimates are required. All households existing during the first wave of interviews (February through April 1979) are considered the initial population. Based on the rules adopted for the following individuals who move, we

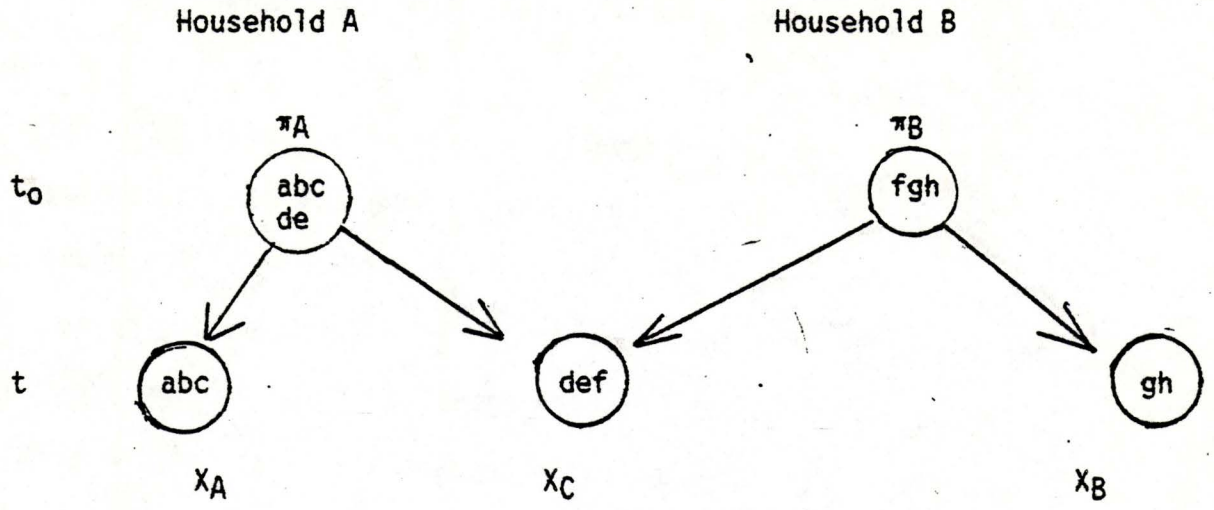
have essentially a longitudinal sample of persons as well as households for the initial population. Since no new sample was drawn at any subsequent interview, the sample does not completely represent the non-institutionalized U.S. population after first quarter of interview. There were persons in the following categories at the initial interview time but became part of the non-institutional population at a subsequent wave of interviewing: 1) U.S. citizens living abroad, 2) citizens of other countries who subsequently move to the U.S., 3) persons in institutions or armed forces barracks. These persons will be called the group R subpopulation which did not have chance to be selected as original sample persons. At a subsequent wave of interviews, the longitudinal sample did not include any household in which all current members were in the group R subpopulation. However, persons in the group R subpopulation who later joined households that included original persons eligible for sampling in the first wave were added to the cross-sectional universe. These person along with new borns will be called "additions" in subsequent waves. In general, "additions" are defined as persons moving into eligible households after the first wave who were not eligible for sampling in the first wave.

III. GENERAL CONCEPT OF CROSS-SECTIONAL ESTIMATION

Due to the procedures adopted for following movers in the 1979 Research Panel, at subsequent interviews a household could consist of members from more than one household in the universe at the time of the first

wave. The inclusion probability of such a household would depend on the inclusion probabilities of the households which the members of the current household were part of at the time of the first interview. The inverse of this inclusion probability is usually used as the weight of such a household in estimation. However, because of the sample design of the 1979 Research Panel, the inclusion probability of a household is a function of its primary sampling unit, type of sampling frames and the 1975 income of the household which occupied the housing unit during the SIE interview. Only the inclusion probability of an original sample household was feasible to calculate. The inclusion probability of an original nonsample household is almost impossible to evaluate. Therefore, some alternative weighting procedures needed to be explored.

The idea to be presented in this discussion is very simple. We will associate observations at any given point in time with the know inclusion probabilities of the original sample households. We will split up observations belonging to a household when current household members come from more than one original household. A portion of the observation is then associated with each original household. The following example will illustrate the idea: Assume that A & B are two original households with inclusion probabilities π_A and π_B respectively. At the first wave of interviews, household A consists of five members, a,b,c,d, and e, and the household B consists of three members, f,g, and h. During the second wave of interviews we find that d,e, and f are living together and form a new household, called household C, while a,b, and c are still in household A and g and h are still in household B.



Two alternatives are proposed, both involving the division of household C into two parts; one part is associated with household A and the other with household B:

a) Multiplicity Approach:

Based on the number of ways (called multiplicity) that the new household C can be included in the sample, the observation (additive, such as counts, income or values) of household C (called X_C) is divided by the number of original households involved (two in this case) and each portion is added to the corresponding observation of household A (called X_A) and household B (called X_B). Therefore, if both households A and B are original sample households, the cross-sectional estimate, $\hat{X}$, for the total at the second wave based on these three households can be expressed as:

$$\hat{X} = \frac{1}{\pi_A}\left(X_A + \frac{1}{2}X_C\right) + \frac{1}{\pi_B}\left(X_B + \frac{1}{2}X_C\right)$$

$$= \frac{1}{\pi_A}X_A + \frac{1}{\pi_B}X_B + \left(\frac{1}{2\pi_A} + \frac{1}{2\pi_B}\right)X_C.$$

Hence, the weight for the new household c is $\frac{1}{2\pi_A} + \frac{1}{2\pi_B}$. If only household A is a sample household, then the weight for the new household is $\frac{1}{2\pi_A}$; if only household B is a sample household then the weight for the new household is $\frac{1}{2\pi_B}$.

b) Fair Share Approach

This approach assumes that all household members contribute equally to their household. Thus, the observation of household C is divided into appropriate portions based on the proportion of members of household C which come from each original household (2/3 from household A and 1/3 from household B in this example). Therefore, if both households A and B are original sample households, the cross-sectional estimate for the total at the second wave based on these three households is expressed as

$$\hat{X} = \frac{1}{\pi_A}\left(X_A + \frac{2}{3}X_C\right) + \frac{1}{\pi_B}\left(X_B + \frac{1}{3}X_C\right)$$

$$= \frac{1}{\pi_A}X_A + \frac{1}{\pi_B}X_B + \left(\frac{2}{3\pi_A} + \frac{1}{3\pi_B}\right)X_C.$$

Hence the weight for the new household C is $\frac{2}{3\pi_A} + \frac{1}{3\pi_B}$. If only household A is a sample household, then the weight for the new household is $\frac{2}{3\pi_A}$; if only household B is a sample household, then the weight for the new household is $\frac{1}{3\pi_B}$.

Since our sample was longitudinal in nature, if the universe remained constant through time, original sample persons would be a representative sample of the universe at any given point in time. Hence, using the inclusion probabilities of the original sample persons, the above estimators are unbiased (proof is given in next section). However, our feasible target population (excluding the group R subpopulation) changes through time by including the 'additions' (defined in Section II). To compensate for this, we will include these "additions" in the proposed estimators below.

IV. PROPOSED ESTIMATORS FOR GENERAL POPULATION (AREA) FRAME

Before the estimators are given, some notation should be defined. For the first wave of interview (time t_o), let

$X(t_o) = \sum_{k=1}^{N(t_o)} X_k(t_o)$ the parameter to be estimated, where $X_k(t_o)$ is the value of the characteristic for the k^{th} unit (which may be a person or a household) and $N(t_o)$ is the number of units at time t_o;

$\alpha_k = 1$ if unit k was in the sample at time t_o, k = 1, $N(t_o)$

$= 0$ otherwise

π_k = the probability that unit k was selected for the sample at the first wave of interview (time t_o)

$= P_r(\alpha_k = 1) = E(\alpha_k)$, k=1, $N(t_o)$

At a subsequent wave (time t), define for each household i:

S_i = the total number of current residents of household i at time t

r_i = the number of original eligible households from which the current residents come (does not include households from which "additions" come)

and

$S_{i1}, S_{i2}, \ldots, S_{ir_i}$ be the number of current residents from each of r_i original households and S_{ia} be the number of current residents from the "additions" as defined in Section II. Write

$$S_i = \sum_{j=1}^{r_i} S_{ij} + S_{ia}$$

$$S_{io} = \sum_{j=1}^{r_i} S_{ij}$$

Also define N(t) as the total number of units in the target population at time t, such as household units (include all the original households plus newly formed households) or other units based on a group of persons such as families or sub-families (include all sample persons, interviewed nonsample persons plus "additions"). And let

$$X(t) = \sum_{i=1}^{N(t)} X_i(t)$$

be the parameter (total) to be estimated for the target population at time t. To simplify the notation, we will replace N(t), X(t) and $X_i(t)$ by N, X and X_i respectively.

Based on the general concept described in Section III, two cross-sectional estimators are proposed for the area frame to estimate the total of a characteristic of the target population at time t.

Estimator I (Multiplicity Estimator):

This estimator is based on the multiplicity of each current household

$$\hat{X}_M = \sum_{i=1}^{N} W_{Mi} X_i$$

where

$$W_{Mi} = \sum_{j=1}^{r_i} \frac{\alpha_j}{r_i \pi_j}.$$

Note that α_j and π_j are associated with original households but are reindexed within each current household i. It is easily seen that

$$E(\hat{X}_M) = E\left(\sum_{i=1}^{N} W_{mi} X_i\right) = E\left(\sum_{i=1}^{N} \sum_{j=1}^{r_i} \frac{\alpha_i}{r_i \pi_j} X_i\right)$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{r_i} \frac{E(\alpha_i)}{r_i \pi_j} X_i = \sum_{i=1}^{N} \frac{1}{r_i} \left(\sum_{j=1}^{r_i} \frac{\pi_j}{\pi_j}\right) X_i = \sum_{i=1}^{N} X_i = X$$

Therefore $\hat{X}_m$ is an unbiased estimator of X.

Estimator II (Fair Share Estimator):

This estimator is motivated by the assumption that all current household members contribute equally to the household in which they reside for the major household characteristic values, such as earnings and welfare benefits.

$$\hat{X}_F = \sum_{i=1}^{N} W_{Fi} X_i$$

where

$$W_{Fi} = \sum_{j=1}^{r_i} \frac{S_{ij}\,\alpha_j}{S_{io}\,\pi_j} .$$

As in the multiplicity estimator, α_j and π_j are associated with original households but are reindexed within each current household i. One can see that $\hat{X}_F$ is also an unbiased estimator of X as follows:

$$E(\hat{X}_F) = E\left(\sum_{i=1}^{N} W_{Fi} X_i \right) = E\left(\sum_{i=1}^{N} \sum_{j=1}^{r_i} \frac{S_{ij}\,\alpha_j}{S_{io}\,\pi_j} X_i \right)$$

$$= \sum_{i=1}^{N} \frac{1}{S_{io}} \left(\sum_{j=1}^{r_i} \frac{S_{ij} E(\alpha_j)}{\pi_j} \right) X_i = \sum_{i=1}^{N} \frac{1}{S_{io}} \left(\sum_{j=1}^{r_i} \frac{S_{ij} E(\alpha_j)}{\pi_j} \right) X_i$$

$$= \sum_{i=1}^{N} X_i = X$$

Note that if household j was not in sample at time t_o, it is unnecessary to know the number of current residents from original household j, S_{ij}, in $\hat{X}_F$ since $\alpha_j = 0$. Also note that because "additions" are not included in the weight calculations, they must be identified and excluded from using either estimator.

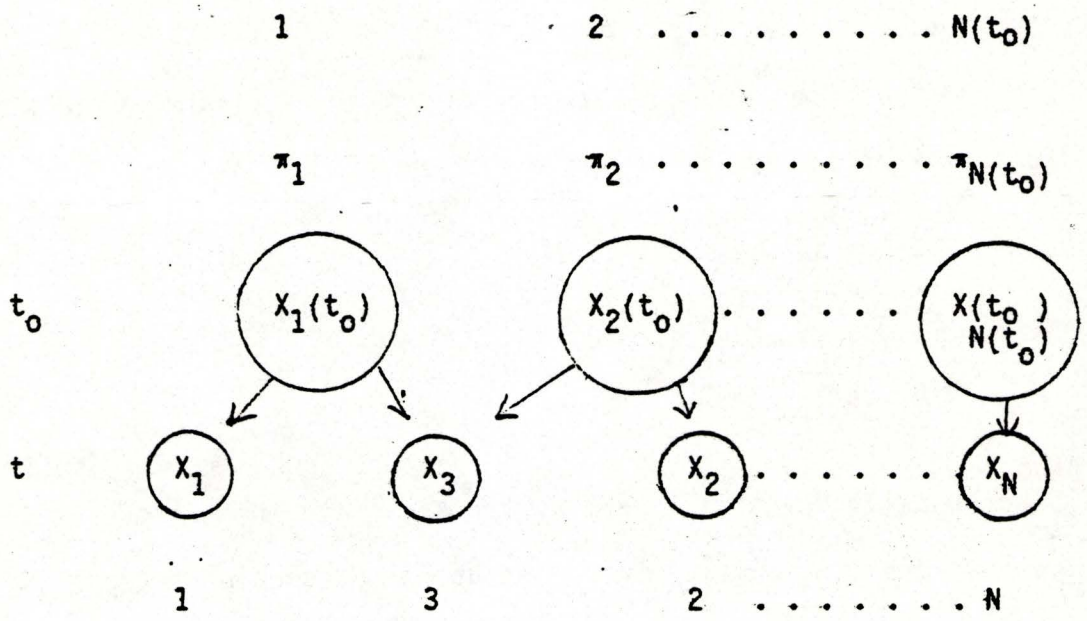
Comparison of Estimator I and Estimator II

Both Estimator I, $\hat{X}_M$, and Estimator II, $\hat{X}_F$, are feasible to compute. We now compare them in both areas of operations and reliability.

In order to compute $\hat{X}_M$, the number of original households eligible for sampling from which the current residents come is needed. This information is particularly difficult to obtain at each successive wave of the survey. However, to compute $\hat{X}_F$ one only needs to know the number of current residents in the household (excluding new additions) and the number or residents from each original sample household. This information could be obtained from the 1979 Research Panel person identifier without collecting additional information.

The equal contribution from the members of a household is a natural assumption. It reflects better the actual share among the household members in the absence of knowledge of the actual contribution from each member. For example, without knowledge of each person's age, employment status and other needed information, it is more logical to assume that earnings and welfare benefits are equally contributed by household members than any arbitrary way of defining household members' shares. And as will be seen below, this assumption also leads to the result that the estimator $\hat{X}_F$ has smaller variance than $\hat{X}_M$.

Assume that at a subsequent wave time t three households are generated from two original households of the first wave of interview (time t_o) as follows:



Let $X_k(t_o)$, k=1, $N(t_o)$ be the value of the characteristics of interest for household k at time t_o and X_j, j=1, N be that value for household j at time t. Using Section III we divide up X_3 in two parts, fX_3 and $(1-f)X_3$ and then associate fX_3 with X_1 and $(1-f)$ X_3 with X_2. Without loss of generality, assume that a sample size of 1 was selected at the first wave, to, with probability π_k, k=1, $N(t_o)$. An unbiased estimator of total, X, at time t can be written as

$$\hat{X} = \frac{\alpha_1}{\pi_1} X_1 + \frac{\alpha_2}{\pi_2} X_2 + \left(\frac{f\alpha_1}{\pi_i} + \frac{(1-f)\alpha_2}{\pi_2} \right) X_3 + \ldots\ldots\ldots\ldots$$

where α_i, i=1, $N(t_o)$is defined at the beginning of this section. Notice that both $\hat{X}_M$ and $\hat{X}_F$ are special cases of $\hat{X}$. The variance of $\hat{X}$ is

$$Var(\hat{X}) = \pi_1 \left\{\left(\frac{1}{\pi_1} X_1 + \frac{f}{\pi_1} X_3\right) - X\right\}^2 + \pi_2 \left\{\left(\frac{1}{\pi_2} X_2 + \frac{1-f}{\pi_2} X_3\right) - X\right\}^2 + \ldots\ldots$$

The remaining terms are not explicitly given here since they are not functions of f. The $Var(\hat{X})$ is minimized if

$$f = \frac{\frac{X_2+X_3}{\pi_2} - \frac{X_1}{\pi_1}}{(\frac{1}{\pi_1} + \frac{1}{\pi_2})X_3}$$

Since usually not both π_1 and π_2 are known and in most of the surveys conducted by the Bureau of the Census, the inclusion probabilities, π_k, are about the same for all ultimate sampling units (even though they are unequal in the 1979 ISDP), one many simplify f to

$$f = \frac{X_2+X_3 - X_1}{2X_3}$$

Obviously, a weight defined as a function of survey observations is not easy to implement. To further simplify f, we assume the growth of X from t_o to t is constant for all units and define

$$a\ X_1(t_o) = X_1 + X_{31}$$
$$a\ X_2(t_o) = X_2 + X_{32}$$
$$X_3 = X_{31} + X_{32}$$

where X_{3i} is the share of X_3 belonging to household members from original household i, i=1,2. Then

$$f = \frac{2\,X_{31} + a(X_2(t_o) - X_1(t_o))}{2(X_{31} + X_{32})}$$

Without knowledge of both $X_1(t_o)$ and $X_2(t_o)$, one would naturally assume that the two initial households are about the same i.e., $X_1(t_o) = X_2(t_o)$ and reduce f to

$$f = \frac{X_{31}}{X_{31} + X_{32}}$$

Now if the contribution to X_3 is proportional to the number of persons from each original household, then

$$f = \frac{S_{31}}{S_{30}}$$

as defined in W_{F1}. This result can be extended to any sample size as well as that the new household members are from more than two original households. Therefore, without knowledge of the actual contribution from each household member, $Var(\hat{X}_F)$ is smaller than $Var(\hat{X}_M)$ under these general conditions.

V. PROPOSED ESTIMATORS FOR LIST FRAMES

Since persons are the list frame sampling units, we can divide all persons in the general population into three groups based on their relationship with the list frame under consideration.

I) Persons who are included in the list frame (called list frame persons). For the SSI list frame, this group includes all the nonaged recipients of the Federal Supplemental Security Income in December 1978; while for the BEOG list frame, this group includes all the eligible applicants of the Basic Educational Opportunity Grant as of September 1978 for school year 1978-79.

II) Persons who are not included in the list frame but live with a list frame person(s) during the first wave of interview (February through April 1979).

III) Persons who are not included in the list frame nor do they live with a list frame person(s) during the first wave of interview.

Both Group I and II had some chance to be included in the list frame sample, but Group III did not. The original (first quarter) households which consist of Group I and/or Group II persons will be called list frame households. As time went on, some members of Group III moved in and lived with person(s) belonging to Group I or II. Such members of Group III will be 'additions' for the list frame, since they are not initially eligible for sampling in the list frame. Note that the type of persons already described as "additions" for the general population (as defined in Section II) will also be "additions" for the list frame. For the following discussions, we now define two types of "additions" for the list frames: the "additions" that come from Group III will be called "Group III additions"and the type of "additions" as defined for the area frame will be called "area frame addition."

If a list of recipients of a government assistance program is used as a list frame then Group III is usually fairly large. If we construct our estimators the same way as we did for the area frame, we will include a lot of Group III persons in our estimates at time t of subsequent interviews. Consequently, we wouldn't really know what "subpopulation" we were estimating. In our opinion, it is not feasible to define such a subpopulation at time t. Without new sample drawn each wave from the updated list, a proper cross-sectional estimate for a list frame subpopulation at time t is not likely, especially if the turnover rate of the list frame members is high. Therefore, we will restrict our cross-sectional estimates to be based on only the original list frame sample persons (that is, the list frame persons selected for list frame sample plus all the persons who reside with them during the first quarter of interview) and the "area frame additions." In so doing we know that at any time t, the target population we are estimating consists of the original list frame subpopulation (that is Groups I and II) and the type of "additions" as defined in the area frame. Note that the original list frame subpopulation is determined by persons who were on the list at the time of sample selection. They may not be on the list by the time of initial interview.

In the 1979 ISDP panel, a household may have a multiple chance of being selected for the list frame sample if more than one member of the list frame persons live in that household at the first wave of interview.

(Some effort was made to reduce multiple chance of selection for those households in SSI frame.) Therefore, the concept of the base weight for the first wave of interview is no longer trivial.

Similar to the area frame, we define $X(t_o) = \sum_{i=1}^{N^L(t_o)} X_i(t_o)$ as the parameter to be estimated from a list frame sample at time t_o, where $X_i(t_o)$ is the value of the characteristic for the i^{th} unit in the list frame subpopulation, which includes both Group I and II defined at the beginning of this section. Let

α_i = 1 if list frame person i is in the sample,

= 0 otherwise (note that $\alpha_i = 0$ for all non-list frame persons)

π_i = the probability that list frame person i is selected for the list frame sample for the first wave of interview (time t_o)

= $Pr(\alpha_i = 1) = E(\alpha_i)$

β_{oj} = the number of list frame persons (indexed by i) in the j^{th} household at time t_o,

α'_j = 1 if the j^{th} household is in the list frame sample,

= 0 otherwise.

Then the base weight at time t_o for the j^{th} household and its residents is defined as

$$W_{oj} = \sum_{i=1}^{\beta_{oj}} \frac{\alpha_i}{\beta_{oj}\pi_i}$$

where α_i and π_i are associated with list frame persons but are reindexed within household j.

For time t of a subsequent wave, let

β_k = the total number of list frame persons living in the original (time t_o) list frame households which the current residents of the k^{th} household come from.

S_k = the total number of current residents at time t; $S_{k1}, S_{k2}, \ldots, S_{kr_k}$ be the number of current residents in the k^{th} household who come from each of r_k original list frame households; S_{ka} is the number of current residents of the k^{th} household who are from the "area frame additions"; and $S_{k\ III}$ is the number of current residents of the k^{th} household who are from the "Group III additions."

Therefore, $S_k = \sum_{j=1}^{r_k} S_{kj} + S_{k\ III} + S_{ka} = S_{kc} + S_{ka}$.

N^L = the total number of units such as household or family units, in the list frame universe at time t (note again that this includes both "area frame additions" and "Group III additions").

The two cross-sectional estimators for the total of a characteristic of the list frame target population at time t are as follows:

Estimator I (Multiplicity Estimator)

To avoid estimating "Group III additions" we will treat all the current residents from the "Group III additions" as a separate list frame sampling unit. Therefore, in the k^{th} household at time t, there are $\beta_k + U_k$ list frame sampling units, where $U_k = 1$, if some of the current residents in the k^{th} household are from "Group III additions," 0 otherwise. The multiplicity estimator for the list frame population total is given in the following:

$$\hat{X}_M^L = \sum_{k=1}^{N^L} W_{Mk}^L X_k$$

where

$$W_{Mk}^L = \sum_{i=1}^{\beta_k} \frac{\alpha_i}{(\beta_k + U_k)\pi_i}.$$

α_i and π_i are associated with original list frame person but are reindexed within each current household k.

Estimator II (Fair Share Estimator)

Motivated by the assumption that all current residents contribute equally to a household we propose the following list frame estimator:

$$\hat{X}_F^L = \sum_{k=1}^{N^L} W_{Fk}^L X_k$$

where

$$W_{Fk}^L = \sum_{j=1}^{r_k} \frac{S_{kj}\alpha_j'}{S_{kc}} W_{oj};$$

and α_j' and W_{oj} are associated with original household but are reindexed within each current household k.

These two estimators are constructed to estimate the list frame subpopulation excluding the "Group III addition." They are not unbiased estimators in global sense. However, the fair share estimator is unbiased under the assumptions that all current residents contribute equally to a household and a household is treated as a fraction of a household if some of the current residents are from "Group III additions." For example, suppose there are five persons, a,b,f,u and v in household M at time t. Among them, a and b are from original household A, f is from original household B and u and v are from "Group III additions." Furthermore, b and f are list frame persons. Denote X_i, i=a,b,f,u,v for the value of characteristic X of the i^{th} person in the household. Then, the expected value of the fair share estimator of the characteristic X for household M is (3/5) $(X_a+X_b+X_f+X_u+X_v)$, and the corresponding value in our target population is $X_a+X_b+X_f$. Let $X_T = X_a+X_b+X_f+X_u+X_v$, then under the assumption of fair share estimator, $X_a+X_b+X_f = \frac{3}{5} X_T = \frac{3}{5} (X_a+ X_b+X_f+X_u+X_v)$. The expected number of households, for household M is $\frac{3}{5}$. For the multiplicity estimator, the situation is quite different. The expected value of X for household M is $\frac{2}{3} (X_a+X_b+X_f+X_u+X_v)$ and the expected number of households for household M is $\frac{2}{3}$. Therefore, more assumptions will have to be imposed before one can declare that the multiplicity estimator of X is unbiased.

In addition to the unbiasedness described above, $\hat{X}^L_F$ is also preferred for the same reasons (operational and reliability) stated in the area frame. In computing $\hat{X}^L_F$, we need to know β_{oj}, the number of list frame persons in a sample household at the initial interview (time t_o). This information was not difficult to obtain. And at any subsequent wave of interview time t, we needed to know only S_{kc}, the total number of current residents who are not "area frame additions" and S_{kj}, the number of current residents from each original list frame sample household. The latter can be obtained through the person identifier.

However, in order to compute $\hat{X}^L_M$ at time t we would have to ask one complicated question to obtain β_k, the total number of list frame persons living in the original households from which the current residents come. For example, for Supplemental Security Income (SSI) list frame at any subsequent interviews we would need to ask:

> During February through April 1979, how many of your household members were blind or disabled recipients of the Federal Supplemental Security Income in December 1978 who were neither institutionalized nor had representative payees?

It was determined to be extremely difficult to collect this information for the ISDP.

VI. SUMMARY

This research was completed before the first interviews of the 1979 ISDP. These two estimators were constructed based on the specific procedure of following movers in the 1979 ISDP. However, they can be easily modified to apply to other designs and procedures. The fair share estimator was actually used for the 1979 ISDP. It is also being used for the 1984 Survey of Income and Program Participation.

As noted in Section III, the inverse of the inclusion probability of a household at time t is usually used as the weight of that household to obtain unbiased estimator. When a household consists of members from two original households (called households i and j), the inclusion probability of this new household is $\pi_i + \pi_j - \pi_{ij}$, when π_{ij} is the joint selection probability of households i and j at the first wave of interview. This inclusion probability is not only operationally impossible to obtain, its inverse can be also reduced to the weight (W_{Mi}) of multiplicity estimator in most surveys conducted by the Census Bureau. In these surveys, the inclusion probabilities are almost the same for all ultimate sampling units and the joint selection probabilities are generally due to the sampling without replacement within PSU which are generally negligible. Therefore, the fair share estimator not only overcomes the trouble of obtaining such inclusion probabilities it is also a more reliable estimator than such traditional estimators under some general conditions and it is easy to implement.

SMD:HHuang:osh:jeg:LSB#2A

BIBLIOGRAPHY

[1] David, M. (ed), (1983). Technical, Conceptual and Administrative Lessons of the Income Survey Development Program. New York: Social Science Research Council.

[2] Herriot, R. and Kasprzyk, D. (1984). The Survey of Income and Program Participation. To be published in the Proceedings of The Social Statistics Section, American Statistical Associates.

[3] Nelson, D., McMillen, D. and Kasprzyk, D. (June 1984). An Overview of the Survey of Income and Program Participation. SIPP Working Paper Series, Number 84-01.

[4] Survey Documentation for 1979 Research Panel, (1983). Department of Commerce, National Technical Information Service, Springfield Virginia.

[5] U.S. Department of Commerce, Bureau of the Census, (1978). Some initial thoughts on Weighting the 1979 Income Survey Development Program. Memorandum from G. Capps to M. Schooley, November 30, 1978.

[6] U.S. Department of Commerce, Bureau of the Census, (1979). Some Alternative Procedures on Weighting the 1979 Income Survey Development Program (SDP). Memorandum from M. Schooley to G. Shapiro, March 1, 1979.

[7] Ycas, M. and Lininger, C., (1981). The Income Survey Development Program: Design Features and Initial Findings. Social Security Bulletin 44:11, November 1981.

WEIGHTING OF PERSONS FOR

SIPP LONGITUDINAL TABULATIONS

by

David R. Judkins, David L. Hubble,
James A. Dorsch, David B. McMillen
and Lawrence R. Ernst

U.S. Bureau of the Census

For Presentation at the American Statistical Association
Annual Meetings - August 1984

I. INTRODUCTION

Since October of 1983, the Census Bureau has been conducting interviews for a new survey, the Survey of Income and Program Participation (SIPP). The survey will effect long-sought improvements in the measurement of annual income and the complex relationships between income flows, labor force participation, participation in government programs such as welfare, and tax policy. One of the products of the interviewing will be a set of longitudinal records on a probability sample of the population. The subject we address in this paper is the weighting of these longitudinal records so that the data may be analyzed.

We are aware of only two precedents for this weighting. They are the National Medical Care Expenditure Survey (NMCES) and the National Medical Care Utilization and Expenditure Survey (NMCUES). The latter was conducted jointly by the Research Triangle Institute and the National Opinion Research Center [2]. Some work was done on the problem for the Income Survey Development Program (ISDP)[6], but it was not implemented. The techniques used by them are among those under consideration for SIPP. Naturally though, we are also considering some new ideas. These ideas are still in a very preliminary form. We are presenting them here to get early reaction and suggestions from the statistical community.

Our general approach consists of three major steps. The first step is to derive an unbiased weight for each longitudinal record. This is not as straightforward as it seems due to the fact that a slightly different set of people is being interviewed each month. Section III discusses this step.

The second step is to make adjustments for those records that are incomplete. We will use imputation when part of an interview is missing. (See Samuhel's paper in this session [3].) We will also probably use imputation when a

whole interview is missing where the missing interview is bracketed by good interviews. Our research on adjusting for records with more than one missing interview is in too preliminary a stage to report on. (One proposal has been made by Little and David [4].)

The third step is to correct for disproportional representation of demographic types to reduce variance and gain some consistency with the Current Population Survey (CPS). Section IV discusses this step.

Before discussing the weighting, it is essential that we define which of the many possible longitudinal universes is the universe for which estimates are to be provided. Section II deals with this problem.

Finally, we mention some of the important features of the design of SIPP. For more details, the reader is encouraged to first read an overview of the survey [5]. Roughly 20,000 households were interviewed between October 1983 and January 1984, inclusively. That set of interviews constitutes the first wave of the 1984 panel of SIPP. The Census Bureau will try to interview the persons in those households an additional seven or eight times in four-month waves, even if they move. We will also interview any persons who "usually reside" with anyone in the original cross-section for at least one-half of a calendar month. This extra interviewing will only be conducted for the time period that the joint residence is maintained. Only the original cross-section is followed through moves.

II. DEFINING THE LONGITUDINAL UNIVERSE OF PERSONS FOR SIPP

The SIPP universe at the beginning of any panel is persons who are members of the civilian non-institutional population, and members of the military not living in barracks on bases. Defining the longitudinal universe is somewhat more complicated. We begin by defining the possible ways persons can enter

and exit this universe. Next we discuss the relationship between the cross-sectional universes and the longitudinal universe. The third topic of this section addresses the definition of table universes, and a discussion of calculating annual income for persons in the longitudinal universe.

There are two ways persons can enter the SIPP universe: 1) persons can move from overseas (immigrate or return), institutions, or from military barracks; 2) persons can be born to members of the universe.

Similarly, there are two methods of exiting the universe; 1) moving overseas, to an institution, or to military barracks 2) dying. Given these conditions of entering and exiting the universe, and a definition of the initial universe, we can define the universe at any subsequent point in time, and the means by which the universe grows and diminishes over time. The next problem is to make the transition from the cross-sectional universes to a single longitudinal universe.

There are three methods of defining a longitudinal universe: 1) the composition can be fixed at some point in time; 2) the universe may be defined as the union of some set of cross-sectional universes; and 3) the universe may be defined as the intersection of some set of cross-sectional universes.

A longitudinal universe may be defined at a given point in time. For example, we can take the civilian noninstitutional population at the time the sample is drawn, at the midpoint of the panel duration, or at the end of the panel to define the universe of interest. Of course, the time point chosen could be any time point within the duration of the panel. This rather narrow

definition of the universe has an advantage in its simplicity, but also several disadvantages. Dependent on the chosen point in time, this definition produces a strictly declining population, a first increasing and then decreasing population, or a strictly increasing population. In the first case all entrants are excluded from the longitudinal universe, and only exits are allowed to alter the universe. In the second case, entry is allowed and exit is denied until the midpoint, when the situation reverses. In the last case, all those who exit during the panel are excluded from the longitudinal universe and only entries are allowed to alter the universe. In addition, it is difficult to argue why one point or another should be chosen as the point in time to define the universe, and for some purposes you may need a different point than the one originally chosen.

The next two definitions build from the above idea that a universe may be defined at any point during the panel. Let us assume then a set of universes each defined at a different point in time. To further simplify discussion, let us assume a set of twelve monthly universes defined at the midpoint of each month. The two options are to use either the union or the intersection of these sets.

Consider first the union of sets. The union of these monthly universes is all persons who were at some point during the year members of the civilian noninstitutional population. In other words, all members of the target population plus all persons who enter or exit during the year are included in the union of sets definition. This is the most inclusive of the universe definitions offered here, and the one which best captures the dynamic characteristics of the population. Some of the disadvantages of this type of definition will be raised below in the discussion of tabulations and

table universes.

An alternative to the union of sets is the intersection of the set of twelve monthly cross-sectional universes. Here we include in the longitudinal universe only those persons who were members of all of the cross-sectional universes. In other words, only those persons who were members of the civilian noninstitutional population or the special military categories on the fifteenth of each of the twelve months. This definition is even more restricted than the point-in-time definition. This intersection of sets definition produces a static population. That is to say there is no entering or exiting allowed.

Of the three longitudinal definitions offered here, only the union of sets incorporates the dynamic qualities that are inherent in a longitudinal process.

That would seem to make it the logical choice; however, this is also the definition that produces the most complications when tabulating data. These and other problems associated with tables are discussed below.

Consider, for example, a simple table of marital status given in figure 1. Here we are tabulating marital status at the beginning of the year with marital status at the end of the year. Thus, cell 1, 1 consists of those persons who were married at both points in time. Cell 1, 2 consists of those persons who were married at the beginning of the year and separated at the end of the year. Given the union of sets definition, there are not sufficient columns to tabulate all persons. In fact, nearly any universe definition will require the addition of at least one column to this table. That is to say, there is no place in this table for persons who were in

the universe at one point in time, and not in the universe at the other point in time. For the union of sets definition there is a need for both a column and a row for persons not in the universe at time 1 or not in the universe at time 2. For those definitions that allow exiting only a column for persons not in the universe at time 2 is necessary as long as the beginning point of the universe and the table are the same.

Similar problems arise in computing annual income. Aggregating across months is simple, but it is not clear how to compare income amounts for full year and part year persons. That is simply to say that a $6,000 income for 6 months and a $6,000 income for 12 months are not the same.

Figure 1

Marital Status 1984	Marital Status 1985				
	Married	Separated	Divorced	Widowed	Never Married
Married					
Separated					
Divorced					
Widowed					
Never Married					

III. INITIAL WEIGHTING

For SIPP, as for ISDP, a cross-section of the population will be followed for a period of time. Data will also be collected on the people that the original cross-section live with. The original idea was that only the data on the people in the original cross-section would be used in person longitudinal tabulations; the data on the other people would be used only to provide the "household experience" of the original cross-section. We are now reexamining that idea. The data on the other people can be used to better understand the experience of new entrants to the SIPP universe. Furthermore, there are ways to use these data more intensively to gain valuable variance reductions. Unfortunately, these procedures require strong assumptions for unbiasedness. In the following sections, we explore the trade-off. We begin with a general discussion, follow with a study plan on the question and some proposed procedures, and close with some examples on the application of the procedures.

A. Variance Reduction Versus Bias Control.

Let us first define some terms and clarify the type of parameters to be estimated. We divide the sample people into three groups. A person is an original sample person if he/she is a member of the original cross-section. [1/] A person is an associated sample person if he/she was a member of the eligible population at the time the cross-section was selected but happened not to be selected. Anyone else is an additional

1/ A person in original cross-section of households who was 15 years old or older at the time of the first interview is definitely an original sample person. Twelve, thirteen, and fourteen year old children are more difficult to classify. At first, no questionnaires are filled out for them and they are not followed in the rare event of an unaccompanied move. However, after they turn 15, they are treated the same as any other original sample person. We will treat them here as original sample people. Children eleven or younger are not classified at all.

sample person. This last group consists of recent discharges from institutions, new immigrants, and people moving out of military barracks. The type of parameter to be estimated is the frequency of some pattern of labor force participation, program participation, income receipt, etcetera, by demographic characteristics, housing characteristic, geographical unit, educational background, etcetera. A simple example is the frequency of women who were receiving public assistance in January 1984 but were not receiving it in December 1984.

The original idea was to estimate parameters like this one by summing the weights of all original sample persons with the desired characteristics. Data on associated and additional sample people are needed only to classify original sample people with respect to household characteristics; for example, was the original sample person living in a household in which at least one member received social security? Given this scheme, no data are needed on associated or additional sample people for the period that they don't reside with original sample people. Hence, we do not follow associated or additional sample people if they separate from original sample people. Clearly then, the data on associated and additional sample people are frequently incomplete.

Despite this incompleteness, we are now considering ways to squeeze more information out of this data. The first way is to provide estimates for the "union" universe using the data on additional sample people. The second way is to use the data on both types to reduce variances. To

begin the argument for this second use, we first point out that for shorter time periods these data are frequently either complete or nonexistent. (Throughout this section, by complete we mean complete ignoring nonresponse.) This is always true for 1 month periods, usually true for 3 month periods, and frequently true for 12 month periods. For example, suppose that Ruth is an original sample person interviewed in October 1983. In November, she marries Jack, who was in the October SIPP universe. They stay together at least through April 1985. Then Jack is an associated sample person on whom we have complete 1984 data. Alternatively, suppose that Jack was living in a military barracks in October 1983. Then he is an additional sample person on whom we have complete 1984 data. There will obviously be many more cases in these complete categories for 1985 data. Furthermore, there will be many cases where we are only missing one or two months of data.

Intuitively, it seems wasteful to give zero weights to these cases with complete or almost complete data, as originally intended. On the other hand, zero weights must be assigned to the seriously incomplete cases to avoid large-scale imputation. One possible solution is obtained by initally assigning strictly positive weights to all cases, including these that are incomplete due to field procedures, and then treating the incomplete cases as if they were caused by non-response. Imputation would be used for the almost complete cases, and a weighting adjustment would be used for the seriously incomplete cases. Note then that the seriously incomplete cases would have zero weights, while the other cases would have positive weights. If enough data has been collected on the associated and additional sample people to correctly model the probability of this type of nonresponse, then we would still have unbiased estimators.

An example the type of model required is that starting from a given socio-economic stratum, the new economic situation of a male divorcee does not depend on whether he or his ex-spouse was the original sample person. Here we stress that if a person has responded to even a single wave of SIPP, then we have an extraordinary wealth of data available for modeling.

Study Plan

Of course, we will never know for certain whether such a model is correct. There is a risk of biasing the estimators, and as a rule the Bureau is willing to risk biases for decreases in variance only if there is some evidence that the bias squared is substantially less than the variance decrease. Our plan at this time is to quantify for each proposed weighting procedure the frequency of positively weighted incomplete cases by the severity of the incompleteness. We will then have to rely on subjective judgement to determine if the variance decrease from aggressive use of incomplete data is worth the increased risk of bias. The only source for this information is the ISDP. We are currently working on ways to get appropriate tabulations for it.

B. Unbiased Weighting Procedures

Below we present a very simple result that characterizes a general class of unbiased procedures. Reflection on this result quickly helps one to understand that there are infinitely many unbiased procedures. Most of them are totally inappropriate, but it is very possible that better and radically different weighting procedures exist than have yet been conceived.

Let $X = \sum_{i}^{N} x_i$ be the parameter of interest to be estimated where x_i is the value of the characteristic for the i^{th} unit. Let y_i be an unbiased estimator of x_i. Let w_i be a random variable associated with the i^{th} unit such that y_i and w_i are independent and $E(w_i) = 1$.

Then $Y = \sum_{i}^{N} w_i y_i$ is an unbiased estimator of X because

$$E(Y) = E(\sum_{i}^{N} w_i y_i) = \sum_{i}^{N} E(w_i)E(y_i) = \sum_{i}^{N} x_i = X.$$

If we had complete response by all units, we could take $y_i = x_i$ and

$$w_i = \begin{cases} \text{inverse probability of selection if } i^{th} \text{ unit is in sample;} \\ 0 \text{ otherwise.} \end{cases}$$

Quite frequently, however, we will have incomplete response and will take y_i to be some imputation. Note that we may have $w_i = 0$ for units that are in sample, but that this may be disadvantageous because it wastes data.

Before we present the unbiased weighting procedures, let us define a term: cross-sectional person weight. The cross-sectional weight for a person is the cross-sectional weight of the household, of which he/she is a member. So, by defining the cross-sectional household weight we are implicitly defining the cross-sectional person weight. For simplicity, assume that the first wave cross-sectional weight for a sample household is the reciprocal of the probability of selection. For all nonsample households in the universe, this weight is zero.

Some household compositions may change during the period between two waves. For these sample households, new weights may have to be calculated to account for the changes. Using a multiplicity estimator of the general type suggested by Sirken [7], the cross-sectional house-

hold weight for any month after the first wave is defined to be the mean of the first wave cross-sectional household weights for all original and associated sample persons residing in the household that month.

In this section we present four longitudinal weighting procedures for computing unbiased estimates for persons. They are all presented in terms of the "union" universe, but they can be easily modified for the "intersection" universe by assigning a zero weight to any person who is not in every one of the 12 cross-sectional universes. In Section III.C we compare the procedures with respect to the use of data collected on associated sample persons and additional sample persons.

Procedure 1. Entry Date Weight (ED)

Each person receives a single longitudinal weight for any time interval that contains at least part of the period for which the person was in the universe, namely the cross-sectional weight for the person at his/her entry date into the universe. For all original and associated sample persons, the entry date into the universe is the start of the panel, so their longitudinal weights are their Wave 1 cross-sectional weights. For those who enter the universe after Wave 1, (additional sample persons), the longitudinal weight is the cross-sectional weight of the household, of which they are a member, as of the date they enter the universe. If the cross-sectional weight of the household at that date is zero, then the additional sample person's longitudinal weight is zero.

Procedure 2. Beginning Date of Time Interval Weight (BDI)

Each person receives a longitudinal weight valid for all time

Procedure 3. "Mid" Date of the Time Interval Weight (MDI)

This procedure is similar to Procedure 2. Each person receives a longitudinal weight valid for a specific time interval. Persons in the universe at the "mid" date of the time interval are assigned their respective cross-sectional weights at that date. The difference is that instead of the person longitudinal weights being determined at the beginning date of the time interval, these weights are determined at some predesignated date within the time interval. Persons that enter the universe during the time interval but after the mid date are assigned their respective cross-sectional weights as of the date they enter it, as in Procedure 1 and 2. Persons who leave the universe before the "mid" date are assigned their respective cross-sectional weights as of the date they leave it.

Procedure 4. Average Cross-Sectional Weight (ACS)

Each person receives a longitudinal weight valid for a specific time interval. Persons that remain in the universe throughout the interval are assigned the average of their respective monthly cross-sectional weights. Persons that enter or leave the universe are assigned the average of their respective monthly cross-sectional weights for the months they were in the universe during the time interval. Positive weights are assigned to all sample persons. A more formal definition is given below.

Let U_i = number of months the i^{th} person was in the universe during the specified time interval

Let C_i = sum of the monthly cross-sectional weights of the i^{th} person in the specified time interval

Then the person longitudinal weight is C_i/U_i.

C. Comparison of Procedures

In this section we describe in detail the types of complete and incomplete cases that are used by each procedure. First, we need to define some notation. Let

t_B = the first month that a person is in the universe,
t_E = the last month that a person is in the universe,

t_1 = the first month that a person is in sample,
t_2 = the last month that a person is in sample,

t_m = the mid-month of the interval of interest.

The description is given in Table 1. The first 14 cases comprise the "intersection" universe. The remaining 32 cases fill out the "union" universe. Each case is marked as having complete, partial or no data for the interval of interest. Of course, all of this is assuming perfect response. The only type of missingness that we are discussing here is that caused by operational procedures. On the right, there is a column for each procedure with an "X" if the procedure uses the case.

The entry date procedure uses the perfect cases 1,15,17, and 18, but does not use the perfect cases 2 and 16; the partial cases 3,5, and 19-27; and cases 12 and 44 for which no relevant data exists. The beginning date of interval and mid date of interval procedures both use all of the perfect cases, more of the partial cases and none of the completely missing cases. We thus think that these two procedures will tend to yield smaller variances than the entry date procedure with possibly some small increase in the risk of bias. The average cross-sectional procedure is the most aggressive in utilizing partial data. It uses all the perfect and partial cases and none of the completely missing cases. Also note that it assigns smaller

Table 1. Case Utilization by Procedure

Case	Preceeding Time Interval	Interval of Interest	Succeeding Time Interval	Completeness	Procedure ED	BDI	MDI	ACS
1	$t_B=t_1$		$t_2 \leq t_E$	Perfect	X	X	X	X
2	$t_B<t_1$		$t_2 \leq t_E$	"		X	X	X
3	$t_B=t_1$	$t_m \leq t_2$	t_E	Partial	X	X	X	X
4	$t_B<t_1$	$t_m \leq t_2$	t_E	"		X	X	X
5	$t_B=t_1$	$t_2<t_m$	t_E	"	X	X		X
6	$t_B<t_1$	$t_2<t_m$	t_E	"		X		X
7	t_B	$t_1 \leq t_m$	$t_2 \leq t_E$	"			X	X
8	t_B	$t_1 \leq t_m \leq t_2$	t_E	"			X	X
9	t_B	$t_1 \leq t_2 < t_m$	t_E	"				X
10	t_B	$t_m < t_1 \leq t_2$	t_E	"				X
11	t_B	$t_m < t_1$	$t_2 \leq t_E$	"				X
12	$t_B=t_1 \leq t_2$		t_E	No Data	X			
13	$t_B<t_1 \leq t_2$		t_E	"				
14	t_B		$t_1 \leq t_2 \leq t_E$	"				
15	$t_B=t_1$	$t_2=t_E$		Perfect	X	X	X	X
16	$t_B<t_1$	$t_2=t_E$		"		X	X	X
17		$t_B=t_1$	$t_2 \leq t_E$	"	X	X	X	X
18		$t_B=t_1$ and $t_2=t_E$		"	X	X	X	X
19	$t_B=t_1$	$t_m \leq t_2 < t_E$		Partial	X	X	X	X
20		$t_B=t_1$	$t_2<t_E$	"	X	X	X	X
21		$t_B=t_1$ and $t_m \leq t_2$	t_E	"	X	X	X	X
22		$t_B=t_1$ and $t_m \leq t_2 < t_E$		"	X	X	X	X
23	$t_B=t_1$	$t_2<t_m \leq t_E$		"	X	X		X
24	$t_B=t_1$	$t_2<t_E<t_m$		"	X	X		X
25		$t_B=t_1 \leq t_2 < t_m$	t_E	"	X	X		X
26		$t_B=t_1 \leq t_2 < t_m \leq t_E$		"	X	X		X
27		$t_B=t_1 \leq t_2 < t_E \leq t_m$		"	X	X		X
28	$t_B<t_1$	$t_m \leq t_2 < t_E$		"		X	X	X
29	$t_B<t_1$	$t_2<t_m \leq t_E$		"		X		X
30	$t_B<t_1$	$t_2<t_E \leq t_m$		"		X		X

Case	Preceeding Time Interval	Interval of Interest	Succeeding Time Interval	Completeness	Procedure ED	BDI	MDI	ACS
31	t_B	$t_1 \le t_m \le t_2 < t_E$		Partial			X	X
32	t_B	$t_1 \le t_m$ and $t_2 = t_E$		"			X	X
33		$t_B < t_1 \le t_m$	$t_2 \le t_E$	"			X	X
34		$t_B < t_1 \le t_m \le t_2$	t_E	"			X	X
35		$t_B < t_1 \le t_m \le t_2 \le t_E$		"			X	X
36	t_B	$t_1 \le t_2 < t_m \le t_E$		"				X
37	t_B	$t_1 \le t_2 < t_E \le t_m$		"				X
38	t_B	$t_m < t_1 \le t_2 \le t_E$		"				X
39		$t_B < t_1 \le t_2 < t_m$	t_E	"				X
40		$t_m < t_B < t_1$	$t_2 \le t_E$	"				X
41		$t_m < t_B < t_1 \le t_2$	t_E	"				X
42		$t_B < t_1 \le t_2 < t_E \le t_m$		"				X
43		$t_m \le t_B < t_1 \le t_2 \le t_E$		"				X
44	$t_B = t_1 \le t_2$	t_E		No Data	X			
45	$t_B < t_1 \le t_2$	t_E		"				
46		t_B	$t_1 \le t_2 \le t_E$	"				

weights, in general, to the partial cases than the perfect cases. We think it will tend to yield the smallest variances with the greatest risk of bias.

D. Examples

In these examples a divorced mother, previously living alone, has one of her children (older then 14) and her widowed mother move into her house in December 1983. All three had been in separate households prior to that date. In March 1984, her widowed mother remarries and her new husband, who also had previously been living along, moves into the house at that time. In May 1984, the child leaves the house and moves into an apartment by himself. It is also given that the divorced mother was an original sample person with a weight of 3600, and the child was an original sample person with a weight of 7200, both from rotation group 1 which was interviewed in October 1983. Determine the longitudinal person weight for each of these four persons for the entire year 1984, for each of the procedures, with the following two scenarios:

1. All four persons were in the universe throughout the life of the sample.

2. The same, except now the widowed mother was in an institution until December 1983.

Let

D = divorced mother

C = child

W = widowed mother

H = husband

Entry Date Procedure

Scenario 1.

Since all four people were in the universe for the first wave, the weights are their first wave cross-sectional weights, that is

D = 3,600, C = 7,200, W = 0, H = 0.

Scenario 2.

The same, except the widowed mother's weight is now the cross-sectional weight of the household in which she was residing when she entered the universe in December, 1983, that is W = 5,400 = (3600+7200)/2.

For the other three procedures, we first compute the cross-sectional household weights associated with each person for every month of 1984, for both scenarios. The results are given in Table 2.

Table 2

Monthly Cross-Sectional Weights

	Scenario 1			Scenario 2		
	Jan.-Feb.	Mar.-Apr.	May-Dec.	Jan.-Feb.	Mar.-Apr.	May-Dec.
D	3,600	2,700	1,200	5,400	3,600	1,800
C	3,600	2,700	7,200	5,400	3,600	7,200
W	3,600	2,700	1,200	5,400	3,600	1,800
H	0	2,700	1,200	0	3,600	1,800

The reason for the differences in the weights between these two scenarios is that the widowed mother does not enter into the denominator in Scenario 2.

From the above table, the weights for the beginning date of time interval, and mid-date of time interval procedures immediately follow, while for the average cross-sectional weight procedure, we simply average over the twelve months. The results are given in Table 3. The weights from the entry date procedure are also shown.

Table 3

Longitudinal Weights

Procedure / Person	Scenario 1				Scenario 2			
	ED	BDI	MDI	ACS	ED	BDI	MDI	ACS
D	3,600	3,600	1,200	1,850	3,600	5,400	1,800	2,700
C	7,200	3,600	7,200	5,850	7,200	5,400	7,200	6,300
W	0	3,600	1,200	1,850	5,400	5,400	1,800	2,700
H	0	0	1,200	1,250	0	0	1,800	1,800

IV. CONTROLS

We are currently considering the adjustment of SIPP longitudinal weights so as to achieve the variance reductions associated with ratio estimation while also causing agreement with SIPP cross-sectional controls on a monthly basis; i.e., in addition to simple undercoverage adjustments we are considering the possibility of forcing the sum of the longitudinal weights of all persons in the universe in a given month to equal the cross-sectional population control for that month. Since longitudinal weights are fixed over time while the universe fluctuates over time, such agreement will not occur unless proper steps are taken to ensure it. We are also considering adjustments to force spouses to have equal longitudinal weights. We are considering these two possibilities in order to enhance the face validity of the survey at the least possible cost of reduced precision.

A. Objectives

The primary reason for ratio adjustment of longitudinal weights is to reduce variances of longitudinal weights by ensuring representativeness with respect to demographic variables which are highly correlated with

the variables to be measured. (This is frequently referred to as post-stratification.) To the extent that it corrects for differential under-coverage, it is also hoped that bias is reduced by ratio adjustment.

A reasonably good adjustment is to proportionately adjust the weights of persons by demographic type in a specified month so that the weighted counts agree with independent population estimates by demographic type for that month. Persons not in sample in the chosen month are assigned the factor for their demographic type. This approach operates under the assumption that the degree to which the sample represents each demographic type is not highly variable over time. This adjustment does not adjust weights to monthly controls other than those for the chosen month. Another approach is to make the adjustment for all persons for each of the 12 data months, then assign to a person the average of the 12 factors for his/her cell. Such an adjustment would tend to be influenced less by the vagaries of sample selection. Addressed here is the more complex problem of adjusting weights for disproportional representation in a manner such that consistency with cross-sectional controls is achieved for each month. This problem has a multitude of solutions. However, the solution we seek should be the one which provides the greatest variance reduction. One possible solution is to first adjust weights as outlined in the above paragraph, then further adjust them so that the desired monthly consistency is achieved while minimizing the amount by which weights are further adjusted. This approach preserves the benefits of the initial adjustment by demographic variables provided that this second adjustment causes relatively small changes in weights.

A further refinement would be to adjust so that spouses have equal weights. Naturally, persons undergo changes in marital status during the year; some persons may have more than one spouse over a one year period. Define a "marriage group" to be a group of persons in the SIPP sample, each of whom has been or is married to at least one other person in the group during the data year. It is possible to perform an adjustment so that all persons in a given marriage group have equal weights. This last adjustment would cause slight disagreements between longitudinal population estimates and monthly controls; we believe that such disagreements could be made arbitrarily small.

B. Outline of Adjustment Process

The basic steps in the adjustment process are as follows:

1. Post-stratification

The demographic types that we are most likely to use in post-stratification are those defined by age, race, sex, and household relationship. These are similar to the types used in post stratification for cross-sectional estimation [1].

Within each rotation of the SIPP sample, all persons would receive an adjustment factor ensuring representativeness of the types discussed above. Two possible adjustments are currently under consideration.

a. One month adjustment

For each rotation of the SIPP sample, an initial adjustment is performed for a single month. The weights of all sample persons in the rotation are adjusted so that, for each cell, (demographic type) the sum of that rotation's person longitudinal weights (through this stage of adjustment) is equal to 1/4

of the cell's cross-sectional control for the chosen month. (The factor 1/4 reflects the fact that only 1/4 of the SIPP sample is designated for interview in a given month.) There may be reasons to choose a particular month for the adjustment, but the month chosen has no effect on this development.

Having computed a post-stratification adjustment factor for each of the above defined cells, the factor for the appropriate cell is also applied to each person who is in the SIPP sample during some part of the data year but not in sample during the month for which the adjustment factors are computed.

b. Multiple month adjustment

A second approach is to compute adjustment factors as in a. for all 12 months of the data year, then average the 12 factors for each cell. Either a weighted or unweighted average might be used. This type of adjustment would tend to be smoother than a one month adjustment, and would likely require less adjustment in 2., immediately below.

2. Cross-sectional Consistency Adjustment

It has been proposed that some form of adjustment be used to cause consistency with cross-sectional controls during each month of the data year. If possible, (i.e., if there are enough sample persons to yield reliable adjustment factors), the adjustment would be performed by some small number of cells (perhaps 4 age x sex cells). For simplicity, a one cell adjustment is discussed here. Since adjustments prior to this one should offer substantial variance reductions, the object of the proposed adjustment would be to achieve

the desired consistency while minimizing the overall adjustment to the current weights. Two approaches to the problem are currently under consideration.

The following notation will be used in both developments. Let

Y_{ij} = the weighted number of persons in the SIPP sample only from month i through month j of the data year, where $i \leq j$. The weight is the adjusted weight from 1. above.

X_{ij} = the (unknown) weighted number of persons analogous to Y_{ij}, after the cross-sectional consistency adjustment.

C_k = the cross-sectional control for the k^{th} month.

There are 78 Y_{ij}'s and 78 X_{ij}'s.

a. Lagrange multiplier approach

This approach seeks to minimize the sum of squared deviations

$$D_1 = \sum_{j=1}^{12} \sum_{i=1}^{j} W_{ij} (Y_{ij} - X_{ij})^2 \qquad (1)$$

subject to the constraints

$$\sum_{i=1}^{k} \sum_{j=k}^{12} X_{ij} = C_k, \; k=1, 2, 3, \ldots, 12, \qquad (2)$$

where W_{ij} is an arbitrar y weight. The problem can be easily solved with Lagrange multipliers.

After solving for X_{ij}, the cross-sectional consistency adjustment factor to be applied to the weight of each person in sample only from the i^{th} through j^{th} months is set equal to X_{ij}/Y_{ij}.

Though this problem generally has a solution, it is possible

that negative or large positive weights will result.

b. Linear programming approach

This approach guarantees non-negative weights but may not provide perfect consistency. The idea here is to minimize

$$D_2 = \sum_{j=1}^{12} \sum_{i=1}^{} |X_{ij} - Y_{ij}| \qquad (3)$$

subject to the constraints (2) and the additional constraints

$$X_{ij} \geq 0, \; i, j = 1, 2, 3, \ldots, 12 \qquad (4)$$

The objective function (3) can be expressed in a form by which the problem can be written as a linear programming problem. This problem, if it has a solution, results in non-negative weights due to the constraints (4). It is possible that the problem has no solution, in which case the constraints must be relaxed to the extent that a solution becomes possible. It is possible to write constraints which keep weights from becoming larger than some arbitrary value; it is possible that these additional constraints could make it necessary to relax other constraints. Alternatively, the constraints might be limited in number by requiring consistency for only a subset of the 12 months.

3. Marriage Group Equalization Adjustment

Recall that a marriage group is defined as a collection of persons, each of whom is or was married to another person in the group during some part of the time period over which longitudinal weights are computed. For consistency purposes, it is desirable that persons within a marriage group have equal longitudinal weights. This can be achieved while ensuring that weights will sum to within some specified tolerance of cross-sectional controls for each month,

using the following iterative procedure.

a. Each person within each marriage group is assigned the average weight of persons in that marriage group, using weights adjusted through the cross-sectional consistency adjustment.

b. One of the procedures (whichever is chosen for cross-sectional adjustment, or perhaps some other method) discussed in 2. is implemented, using weights adjusted for cross-sectional consistency to determine Y_{ij} and defining X_{ij} to be the weighted number of persons, analogous to Y_{ij}, after the current adjustment. This yields an adjustment factor to be applied to each person's weight.

c. Steps like a. and b. above can be carried out, one after another, continuing to use the most current weight. Each time step a. is repeated, a check is made to determine whether all Y_{ij}'s are within the specified tolerance of the respective Y_{ij}'s from 2.

d. When the tolerance is met or exceeded by each Y_{ij}, the procedure is terminated and final longitudinal weights are assigned as follows:

   i. Each person in a marriage group receives the last average weight computed for his/her marriage group.

   ii. Each remaining person is assigned a final longitudinal weight equal to his/her weight, through cross-sectional consistency adjustment, multiplied by the product of the factors computed at each successive operation b. above.

It has not been determined whether the above procedure would

necessarily converge, although convergence appears likely. A final remark is that neither of these last two described adjustments would be completely beneficial. They would cause some (to our belief small) increase in mean square error.

REFERENCES

[1] Census Bureau memorandum from C. Jones to T. Walsh, "Cross-Sectional Weigting Specifications for the First Wave of the 1984 Panel of the Survey of Income and Program Participation (SIPP)," November 25, 1983.

[2] Jones, Bruce L., "Development of Sample Weights for the National Household Survey Component of the National Medical Care Utilization and Medicare Utilization and Expenditure Survey," April 1982.

[3] Samuhel, Michael E., "Longitudinal Item Imputation in a Complex Survey," presented to the Survey Research Methods Section of the American Statistical Association during the 1984 Annual Meetings.

[4] Little, R.J.A. and David, M., "Weighting Adjustments for Nonresponse in Panel Surveys," 1983 Working Paper.

[5] Nelson, D., McMillen, D., and Kasprzyk, D., "An Overview of the Survey of Income and Program Participation," SIPP Working Paper Series No. 8401. U.S. Bureau of the Census, Washington, D.C. 1984.

[6] Kasprzyk, D. and Kalton, G., "Longitudinal Weighting in the Income Survey Development Program," in Technical, Conceptual and Administrative Lessons of the Income Survey Development Program (ISDP), Papers presented at a conference, October 6-7, 1982. Social Science Research Council, Washington D.C., 1983.

[7] Sirken, Monroe G., "Household Surveys with Multiplicity," Journal of the American Statistical Association, 65, No. 329 (1970), 257-66.

LONGITUDINAL FAMILY AND HOUSEHOLD ESTIMATION IN SIPP

by

Lawrence R. Ernst, David L. Hubble, and David R. Judkins
Bureau of the Census

For Presentation at the
American Statistical Association Annual Meeting
Philadelphia, Pennsylvania
August 1984

1. INTRODUCTION

Many types of statistics will be produced by the Survey of Income and Program Participation (SIPP), but there is one type that was the driving force behind the unique design of the survey. To be fully successful, SIPP must tell us what happens to households over the course of time. From it we must obtain estimates of the patterns of income receipt, program participation, and labor force participation at the household and family level by a host of other characteristics. Of particular interest are parameters such as total annual household income and the number of families that have stopped drawing food stamps by demographic characteristics.

Before estimates can be produced, a decision must be made on the definition of a longitudinal household to be used in this survey. (To simplify the presentation, we will concentrate our discussion on longitudinal households as opposed to longitudinal families. However, parallel longitudinal estimation procedures can readily be developed for families). It often happens that the occupants of several housing units move and regroup. We need to know which, if any, of the resulting households are to be considered continuations of the previous households. Many definitions have been proposed, but final agreement has thus far not been achieved. Also decisions have yet to be made on whether households that form or dissolve during a time interval of interest are to be considered as part of the universe for estimation purposes. Because of the absence of agreement in these areas, several proposed definition and universe combinations will be considered in this paper. They are listed in Section 2. Also because of this absence of agreement, the major aim of this paper will be simply to compare several possible longitudinal household estimation procedures and present criteria for choosing among them, without attempting to reach a conclusion on a preferred procedure.

We foresee several steps in the process of producing longitudinal households estimates. The focus in this paper, except for the final section, is the first step, the production of weights that would yield unbiased estimates assuming there are no data that are missing or in error, and that the frame coverage is perfect. Several procedures for obtaining such weights will be presented in Section 3. In Section 4 some numerical examples of the weights produced by these procedures are given. Choosing among these procedures is complicated by the fact that even assuming perfect response, data needed to produce unbiased estimates will be missing for some households because they are not collected with the current field procedures. This difficulty is principally due to the fact that, except for a few household definitions, all unbiased procedures assign positive weights to some longitudinal households for time periods when they are not in sample. The severity of this problem and the extent to which it is correctable in the future by changing field procedures or by modeling the missing data, vary by procedure. This problem, along with descriptions of other important features, both positive and negative, that estimation procedures may possess is presented in Section 5. This is followed in Section 6 by a detailed comparison of the features of the estimation procedures under consideration in this paper. Finally, in Section 7 we briefly discuss adjustments to the unbiased weights. It is anticipated that the two major components of such adjustments will be a procedure for adjusting for missing data, and a method for controlling key variables to independent estimates, such as CPS estimates.

It is assumed in this paper the reader has a basic knowledge of SIPP, including the design of this survey. Nelson, McMillen, and Kasprzyk (1984) provides this information.

2. LONGITUDINAL HOUSEHOLD DEFINITIONS

In this section four possible longitudinal household definitions are presented to illustrate the longitudinal weighting procedures that will be described in the next section. A thorough discussion of longitudinal household definitions is presented in McMillen and Herriot (1984). In addition, several other terms will be defined, including the longitudinal household universes considered in this paper.

Since household composition and data for SIPP are obtained on a monthly basis, each of the definitions to be presented will be in terms of household continuity from one month to the following month. A longitudinal household over a time interval of n (>2) months is then defined to be one which is continuous for each of the n-1 corresponding pairs of consecutive months. (It has not yet been decided if this approach will actually be used in SIPP.)

For each of the definitions below the conditions for which household B at month t+1 is the continuation of household A at month t are stated. One condition that we require that all the definitions share is that A and B are either both family households or both non-family households. The other conditions are:

No Change Definition (NC). A and B have the same household members.

Same Householder (SH). A and B have the same householder. As an alternative, householder could be replaced by principal person in this definition without altering any of the statements made about it in subsequent sections, provided the final estimation procedure in Section 3 is also modified accordingly. (The householder of a household is, roughly, the person who owns or rents the housing unit. The principal person is the wife in a married-couple household, and the householder in all other households.)

Reciprocal Majority (RM). The majority of individuals who are both household members of A at time t and in the universe at time t+1 are members

of B at time t+1, and the majority of individuals who are both household members of B at time t+1 and in the universe at time t are members of A at time t. (This type of longitudinal definition was originally developed by Dicker and Casady (1982) for use in the National Medical Care Utilization and Expenditure Survey (NMCUES).)

Shared Experiences Definition (SE). Either conditions (1.a, b) or (2.a-e) presented below are satisfied.

(1.a) A and B are nonfamily households with the same householder, including single person households.

(b) At least 1/2 the members of A are members of B.

(2.a) A and B are family households.

(b) The householder or spouse of the householder of A is the householder or spouse of the householder of B.

(c) A and B have at least two members in common.

(d) If another household A' when substituted for A in (2.a-c) satisfies these conditions, then the number of household members common to A and B is more than the number common to A' and B.

(e) If another household B' when substituted for B in (2.a-c) satisfies these conditions, then the number of household members common to A and B is more than the number common to A and B'.

Some variation of this last definition is currently the leading candidate to be chosen as the SIPP longitudinal household definition.

We will now clarify several other terms.

A household is said to be in existence over a time interval of $n \geq 2$ months if it is longitudinal over that time interval. Its period of existence is the longest such time interval. In the case of a household which is defined cross-sectionally for a month t, but is not longitudinal over either of the two

month intervals containing t, then the period of existence of the household is defined to be one month.

If t_1 and t_2 are any pair of months, and longitudinal estimates are to be made over the interval $[t_1, t_2]$, then the following two possibilities will be considered in subsequent sections for the universe of households for which estimates will be produced.

Restricted Universe. The set of all households in existence over the entire interval $[t_1, t_2]$.

Unrestricted Universe. The set of all household in existence for one or more months in $[t_1, t_2]$.

Each sample panel is interviewed eight times. Each of the eight rounds of interviews takes four consecutive months to complete and is known as a wave.

Finally, we define an original sample person to be a person that was in sample during the first wave and will be at least 15 years of age by the end of the panel.

3. UNBIASED WEIGHTING PROCEDURES

In this section we present five weighting procedures for computing estimates of totals or proportions for longitudinal households that would be unbiased in the sense that the expected value of the estimator over all possible samples is the parameter of interest assuming no data are missing or in error, and perfect frame coverage. Modifications and adjustments of these estimation procedures necessary because of the unrealistic nature of these assumptions will be considered in Section 7. Except for the Continuous Household Members procedure, which will only be applied to the restricted universe, all the procedures will be stated for the unrestricted universe. To apply them to the restricted universe simply zero weight each household which is not in continuous existence over the time interval of interest.

Furthermore, unless otherwise stated, all the procedures will be applied to all four longitudinal definitions defined in Section 2.

First we will explain why a common method of estimation, weighting by the reciprocal of the probability of selection is not feasible for our purposes, and hence the need to consider alternative procedures. Let $X = \sum_{i=1}^{N} x_i$ be a parameter of interest, where x_i is the value of the characteristic for i-th unit in a population of size N. Typically in survey work, to estimate X a sample would be drawn in such a manner that the i-th unit has a known positive probability p_i of being chosen, and X would then be estimated by

$$\hat{X} = \sum_{i=1}^{N} w_i x_i \ , \tag{3.1}$$

where

$$w_i = \begin{cases} \dfrac{1}{p_i} & \text{if the i-th unit is in sample,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

Unfortunately for household and family estimation in SIPP, both cross-sectionally and longitudinally, such an estimation approach is not practical. For example, cross-sectionally a household is interviewed and used in the estimation process for a given month if and only if at least one household member is an original sample person. Consequently, to use (3.1) and (3.2) as an estimator it would be necessary to determine the probability that at least one member of the current household is an original sample person. It would be operationally impossible to determine this probability, since it would first be necessary to determine the first wave households for all current household members and then compute the probability that at least one of these first wave households was selected.

Fortunately though, it is not necessary that w_i satisfy (3.2) in order that (3.1) be unbiased. In fact if w_i is any random variable associated with the i-th unit in the population satisfying

$$E(w_i) = 1, \tag{3.3}$$

then (3.1) is unbiased, that is $E(\hat{X}) = X$. Thus, defining unbiased longitudinal household and family weighting procedures reduces to defining random variables w_i satisfying (3.3).

Before we present the longitudinal weighting procedures we will state what, for purposes of this paper, a cross-sectional household weight is, since most of longitudinal weighting procedures will be defined in terms of cross-sectional weights. The first wave cross-sectional weight for a sample household is taken here to be the reciprocal of the probability of selection. For all nonsample households in the universe this weight is defined to be zero. For any month after the first wave a different definition is necessary because of possible changes in household composition. So, the cross-sectional household weight for any such month is defined to be the mean of the first wave cross-sectional household weights for all persons in the household that month who will be at least 15 years of age by the end of the panel and who were in the universe during the first wave. This type of weighting procedure is currently being used in SIPP to produce cross-sectional estimates, hence the name. It is readily verifiable that the weights satisfy (3.3).

We also will leave it to the reader to verify that the weights for each of the longitudinal procedures to be presented satisfy (3.3) and hence lead to unbiased estimators.

Beginning Date of Household Procedure (BH). Each longitudinal household receives a single weight valid for any time interval that contains at least part of the period for which the household existed, namely the cross-sectional

weight for the household at the beginning date of the household. In particular, if there were no original sample persons in a household at its beginning date then its longitudinal weight would be zero. This approach to longitudinal household estimation was previously used in the NMCUES (Whitmore, Cox and Folsom 1982).

Beginning Date of Time Interval Procedure (BI). Each longitudinal household receives a longitudinal weight valid for all time intervals with the same beginning date, namely the cross-sectional weight for the household at the beginning date of the time interval. Longitudinal households that form during the time interval are assigned the cross-sectional weight for the household at its beginning date, as in the preceeding procedure.

Continuous Household Members Procedure (CM). The following procedure will only be applied to the restricted universe, as defined in Section 2. For any time interval for which the household is in existence the longitudinal weight to be assigned is determined by the set of persons that are members of the household throughout the time interval. The longitudinal household weight is the cross-sectional weight that would be assigned to a household consisting of this set of persons; that is, the average of the first wave weights of these people. A longitudinal weight of zero is assigned to the household if there are no original sample persons who are members throughout the time interval. The procedure is slightly biased because a longitudinal household with no members continuously present throughout a time interval has no chance of receiving a positive weight, thereby making satisfaction of (3.3) impossible. Since we believe this situation will rarely occur, at least for the longitudinal household definitions considered here, we expect this bias to be very small.

Average Cross-Sectional Household Weight Procedure (AW). Each longitudinal household receives a longitudinal weight valid for a specific time

interval, namely the average of the monthly cross-sectional weights for the household over the intersection of the life of the household and the specified time interval.

Note, there are many procedures, like AW, that entail the averaging of weights, both household cross-sectional weights and person longitudinal weights. We will examine only one of these procedures here, as an example of this type of longitudinal household weighting procedure.

Householder Weight Procedure (HW). The following procedure will be applied only to the No Change and Same Householder Definitions, since it is appropriate only for definitions that allow for a single householder during the household's existence (Generalizations of this procedure which are not so restricted in their applicability exist but will not be considered here.) The procedure assigns a single weight valid for any time interval that contains at least part of the period for which the household existed, namely the first wave cross-sectional household weight of the householder's first wave household. A longitudinal weight of zero is assigned to the household if the householder was not an original sample person.

As will be seen in Section 5, this procedure is clearly the one of choice when the Same Householder Definition is used. If that type of definition is used with householder replaced by principal person then a similar modification of this estimation procedure with householder replaced by principal person would be appropriate.

4. EXAMPLES

In the following examples, the cross-sectional weight for the second and subsequent waves will be as defined in Section 3. The longitudinal household definition for procedures BH, BI, CM, and AW will be the reciprocal majority rule, as given in Section 2. For procedure HW, the longitudinal household definition will be the same householder rule, as given in Section 2.

In these examples a divorced mother (householder) with two children (both older than 14) residing with her has her widowed mother move into her house in December, 1983. In August, 1984 her widowed mother remarries and the new husband moves into the house at that time. In April, 1985 one of the children leaves the household. The longitudinal household weights will be determined for the three procedures for the following time periods:

A. the entire year 1984;

B. the entire year 1985;

C. the entire years 1984-85.

This will be done in each case for the following two scenarios:

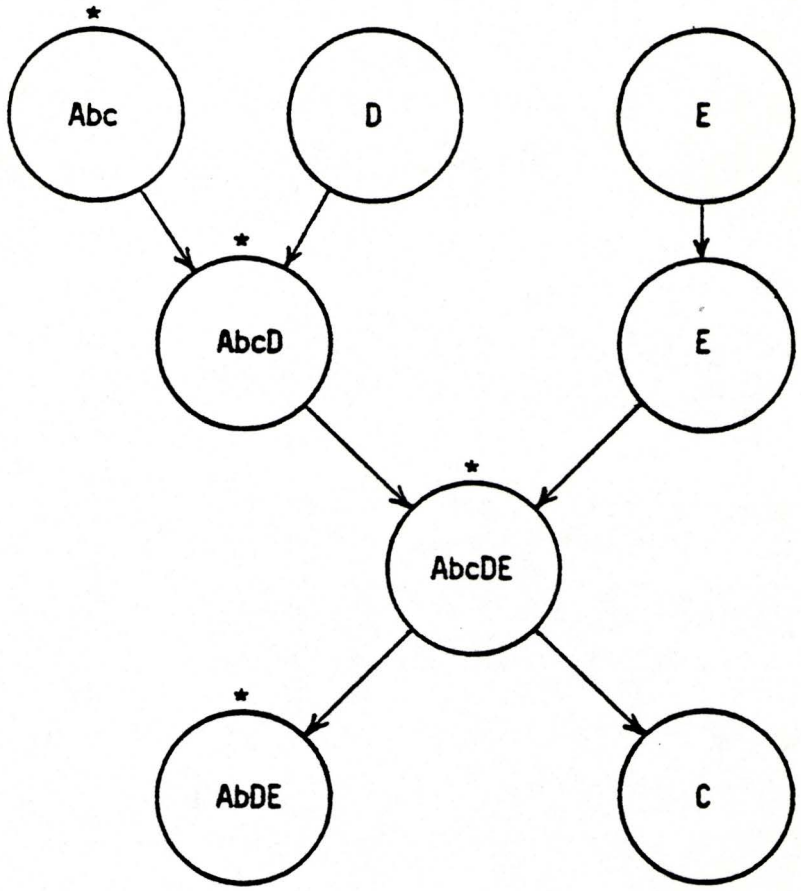
1. the new husband of the widowed mother was the only original sample person in the 1984 SIPP panel (originally interviewed in October, 1983-rotation group 1), with a first wave weight of 8,000;

2. in addition, the divorced mother and her two children were original sample persons (rotation group 1), each with a first wave weight of 4,000.

The six time period, scenario combinations will be denoted by A.1, A.2, B.1, B.2, C.1 and C.2.

Note: We chose to determine the weights only for the longitudinal household that continues through the entire 1984-1985 period, which is marked with an asterisk above it. The other longitudinal households can also be weighted with all these procedures, except CM which applies only to the restricted universe.

Below is a schematic diagram of the example



A = divorced mother

b = c = divorced mother's child

D = divorced mother's widowed mother

E = widowed mother's new husband

Let W_{C1} = cross-sectional weight under scenario 1
W_{C2} = cross-sectional weight under scenario 2

Procedure BH

A.1., B.1., C.1. = W_{C1} for Abc = 0

A.2., B.2., C.2. = W_{C2} for Abc = 4,000

Procedure BI

A.1., C.1. = W_{C1} for AbcD = 0

B.1. = W_{C1} for AbcDE = (1/5) x 8,000 = 1,600

A.2., C.2. = W_{C2} for AbcD = (3/4) x 4,000 = 3,000

B.2. = W_{C2} for AbcDE = (3/5) x 4,000 + (1/5) x 8,000 = 4,000

Procedure CM

A.1. = W_{C1} for AbcD (the continuous members for the time period) = 0

B.1. = W_{C1} for AbDE (the continuous members for the time period)
= (1/4) x 8,000 = 8,000 = 2,000

C.1. = W_{C1} for AbD (the continuous members for the time period) = 0

A.2. = W_{C2} for AbcD (the continuous members for the time period)
(3/4) x 4,000 = 3,000

B.2. = W_{C2} for AbDE (the continuous members for the time period)
= (2/4) x 4,000 + (1/4) x 8,000 = 4,000

C.2. = W_{C2} for AbD (the continuous members for the time period)
= (2/3) x 4,000 = 2,666.67

Procedure AW

A.1. = [[(W_{C1} for AbcD) · 7 months] + [(W_{C1} for AbcDE) · 5 months]]/12 months
= [[(0) · 7] + [(1,600) · 5]]/12 = 666.67

B.1. = [[(W_{C1} for AbcDE) · 3 months] + [(W_{C1} for AbDE) · 9 months]]/12 months
= [[(1,600) · 3] + [(2,000) · 9]]/12 = 1,900

C.1. = [[(W_{C1} for AbcD) · 7 months] + [(W_{C1} for AbcDE) · 8 months] + [(W_{C1} for AbDE) · 9 months]]/24 months

= [[(0) · 7] + [(1,600) · 8] + [(2,000) · 9]]/24 = 1,283.33

A.2. = [[(W_{C2} for AbcD) · 7 months] + [(W_{C2} for AbcDE) · 5 months]]/12 months

= [[(3,000) · 7] + [(4,000) · 5]]/12 = 3,416.67

B.2. = [[(W_{C2} for AbcDE) · 3 months] + [(W_{C2} for AbDE) · 9 months]]/12 months

= [[(4,000) · 3] + [(4,000) · 9]]/12 = 4,000

C.2. = [[(W_{C2} for AbcD) · 7 months] + [(W_{C2} for AbcDE) · 8 months] + [(W_{C2} for AbDE) · 9 months]]/24 months

= [[(3,000) · 7] + [(4,000) · 8] + [(4,000) · 9]]/24 = 3,708.33

Procedure HW

A.1., B.1., C.1. = first wave cross-sectional weight for A = 0

A.2., B.2., C.2. = first wave cross-sectional weight for A = 4,000

5. POTENTIAL ADVANTAGES AND DISADVANTAGES

The ideal unbiased weighting procedure would provide a single set of weights applicable to any time interval, require no more data than were collected, and possess the minimum variance among all unbiased procedures. Unfortunately, no such procedure exists. The procedures described in Section 3 all fail one or more of these three criteria to various degrees. In this section, we explain the nature of the failures without explicitly comparing the procedures. That is done in Section 5.

Multiplicity of Weights. Some procedures have the advantage of assigning to each household a single weight which depends only on conditions as of the first reference month for the household and which is valid for every interval

that the household is in the universe. Other procedures have the disadvantage of sometimes producing different weights for the same household for different time intervals. (Procedures with this disadvantage could be modified so that only a single weight applies to any time interval, by computing for each household the weight appropriate for that procedure for the unrestricted universe and the 2 1/2 year time interval corresponding to the life of the panel. The weight obtained would also be used for any smaller subinterval for which the household is in the universe. However, weights obtained in this manner might not be able to be determined until the end of the life of the panel. This would make them difficult to use because we would have to wait until the last data from the panel were processed before estimates could be produced for any earlier time period. In any case, such weights would often lead to higher variances for short time intervals than weights developed specifically for the short time intervals.)

Unavailable Data Requirements. Most definition and procedure combinations require data from some households for time periods when the household is in existence but not in sample, that is for time periods for which interviews are not conducted for the household because no original sample people are members of the household. This needed data could be information for determining proper longitudinal weights or subject-matter information for use in tabulating the estimates. Some of this information is not collected for the 1984 panel of SIPP because of the current operational procedures. This is a consequence of the fact that agreement has not been reached on the longitudinal household definition to be used in SIPP. In this vacuum, operational procedures were determined mainly by considerations of difficulty and cost. Once a definition has been agreed on, depending on the nature of the unavailable data, it might be possible to change operational procedures for future SIPP panels so that

the required data are collected. To understand the problem with current operational procedures, consider the following situation. A household is longitudinal from month t_B to t_E. Original sample people are part of the longitudinal household only from month t_1 to t_2. If $t_B<t_1$, then some prior information may be unavailable. Revised operational procedures to obtain this information might involve retrospective questions, longer reference periods or proxy data on anyone who left the household before the first interview. If $t_2<t_E$, then some posterior information may be unavailable. Revised operational procedures might involve interviewing the household through t_E.

One of the important discriminants between the weighting procedures is how successfully they avoid the need for data from the period that the longitudinal household exists but is not in sample. (The need for such data is avoided by assigning zero weights to these problem households.) In terms of information needed for weighting, some procedures require only enough data to determine whether $t_B<t_1$, while others need to know t_B even when it is less than t_1. Similarly, some procedures only require knowledge of whether $t_2<t_E$, while others need to know t_E even when it is greater than t_2. Furthermore, besides this need for information for determination of weights, if any parameters other than the number of longitudinal households are to be estimated, then required subject-matter data may be missing as well, either before t_1, after t_2, or both.

While the problem of missing information is a serious one, it is not fatal. Procedures can be developed to compensate for the unavailable data. Specifically, the data collected on these households while they were in sample should be sufficient for performing imputation for existence/nonexistence outside the in-sample period and formation and/or dissolution dates. The imputed values can then be used to calculate weights for these households. These households can then be treated as noninterviews so that

the weights of mover households with similar demograpahic characteristics but with complete data receive increased weights while the deficient households themselves receives zero weights.

If the models underlying the procedures developed for adjusting for the missing information are true then it is still possible to obtain unbiased estimators, although now in a model-based sense. Furthermore, since the missing information that we are concerned with here is not caused by refusal to respond, modeling in this context might not suffer from the usually imperfect assumptions on similarity between respondents and nonrespondents that underlie any adjustments that use data from respondents to account for data missing from refusals. In addition, because of the longitudinal nature of the survey, there is generally a large amount of data available from the problem households that could be used in such adjustments. However, if the models are not perfect, then in general, the larger the proportion of data required that is unavailable, the greater the potential for serious bias problems.

Variances. In general, estimation procedures with the smallest variances are those that utilize available data intensively and tailor the weights to the specific time interval of interest. Unfortunately, as shall be seen in the next section, such procedures are often characterized by heavy needs for unavailable data which, as noted above, may impact unfavorably upon bias. Thus, there often is a direct trade-off between variance and the risk of bias. It will be difficult to weigh these factors against each other, since it appears that no single procedure will provide the correct balance for all of the multitude of characteristics that will be estimated by SIPP.

For use in the next section, we will define some labels for the advantages and disadvantages identified in the foregoing discussion. Let:

T_1 mean that a single longitudinal weight exists for each household, valid for all time intervals for which the household is in the universe, and which depends only on conditions which could be determined during the first interview,

T_2 mean the negation of T_1,

BW_1 mean that no data from the period preceeding the first interview are unavailable but required for weighting,

BW_2 mean that we need to know for weighting whether the longitudinal household existed before the first interview,

BW_3 mean that we need to know for weighting the conception date of the household (within the time interval of interest),

BD_1 mean that no subject-matter data from the period preceeding the first interview are unavailable but required,

BD_2 mean the negation of BD_1,

FW_1 mean that no data from the period following the last interview are unavailable but required for weighting,

FW_2 mean that we need to know for weighting the dissolution date of the household (within the time interval of interest),

FD_1 mean that no subject-matter data from the period following the last interview are unavailable but required,

FD_2 mean the negation of FD_1.

Note that T_1, BW_1, BD_1, FW_1 and FD_1 are the desirable properties.

6. DETAILED COMPARISONS OF ADVANTAGES AND DISADVANTAGES

Table 1 below presents advantages and disadvantages of each definition procedure and universe combination. A comparison of these features follows the table. Next, an explanation of each entry in the table is given. Finally, a discussion of data utilization, which is not in Table 1, is presented.

Table 1.

Features

Definition	Procedures	Universe	T_1	T_2	BW_1	BW_2	BW_3	BD_1	BD_2	FW_1	FW_2	FD_1	FD_2
No Change (NC)	All	Both	X		X			X		X		X	
Same Householder (SH)	Householder Weight (HW)	Both	X		X			X		X		X	
Same Householder (SH) Reciprocal Majority (RM) Shared Experiences (SE)	Beginning Date of Household (BH)	Unrestricted	X			X		X		X			X
SH, RM, SE	BH	Restricted	X			X		X			X		X
SH, RM, SE	Beginning Date of Time Interval (BI)	Unrestricted		X		X		X		X			X
SH, RM, SE	BI	Restricted		X	X			X			X		X
SH, RM, SE	Continuous Household Members (CM)	Restricted		X	X			X		X		X	
SH, RM, SE	Average Cross-Sectional Weight (AW)	Both		X			X		X		X		X

Comparison of Features in Table 1. As noted at the end of Section 4, T_1, BW_1, BD_1, FW_1, and FD_1 are the desirable properties. For the NC definition all five procedures considered here possess all these desirable properties, as does the HW procedure for the SH definition.

However, for the SH, RM, and SE definitions, and most other definitions too, the BH, BI, and CM procedures have different subsets of the set of desirable features, so that the procedure to be adopted depends, at least in part on the features deemed most important. AW possesses none of these desirable features for there three definitions. Its principal advantage lies in possible reductions in variances because of complete utilization of available data, which will be discussed later. BH has advantages T_1, BD_1, and FW_1 for the unrestricted universe, and T_1 and BD_1 for the restricted universe. The main reason for consideration of this procedure would be that it is the only one among BH, BI and CM that always has advantage T_1. BI has advantages BD_1 and FW_1 for the unrestricted universe and BW_1 and BD_1 for the restricted universe. Its principal advantage over BH is that for the restricted universe no retrospective questions need be asked. CM (which is only applicable to the restricted universe) possesses all desirable features except T_1, that is no information not currently collected is needed for this procedure. Recall, however, that CM had the disadvantage of being slightly biased as explained in Section 3.

Explanation of Entries in Table 1. All explanations presented below apply to both universes unless otherwise stated.

NC Definition, All Procedures. Since the composition of a household is unchanged throughout its period of existence under NC, we have the following two possibilities:

(a) No original sample people were in the household at any time during its period of existence, in which case the longitudinal household weight is zero for any time interval and procedure.

(b) One or more original sample people were in the household throughout its existence, in which case the beginning and ending dates of the household are known, as is the composition of the household and complete data for each month of its existence. Consequently, features BW_1, BD_1, FW_1, and FD_1 apply.

Furthermore, T_1 applies since procedures BH, BI, CM, and AW all reduce to the cross-sectional household weight at the beginning date of the household, while HW is the weight of the householder at the beginning date.

SH Definition, HW Procedure. The explanation is similar to the one given above, except now the two cases are: (a) The householder was not an original sample person. (b) The householder was an original sample person.

SH, RM, and SE Definitions, BH Procedure. T_1 is applicable, since by definition the weight is the cross-sectional household weight as of the beginning date of the household. BW_2 applies because the longitudinal household weight is the cross-sectional household weight as of the first month in sample if the household began that month, while otherwise the weight will be zero since there were no original sample people in the household when it began. (For the restricted universe, households which entered sample after the beginning of the time interval always receive a zero weight.)

BD_1 holds since all households with positive weights were in sample at their beginning date and no retrospective subject-matter data is therefore needed.

FW_1 holds for the unrestricted universe since the weight is determined at the beginning date of the household. However, for the restricted universe, it is necessary to know if the household continued to exist throughout the entire time interval because it receives a zero weight for the time interval if it did not continue. Under current procedures a household which no longer has any original sample person is not followed, and it would therefore generally

not be possible to determine if it remained in existence for the entire time interval. Consequently, FW_2 applies.

FD_2 applies since there would be missing data for all households with positive weights which continued to exist after there were no longer any original sample people present, which could happen for any of there three definitions.

SH, RM, and SE Definitions, BI Procedure. T_2 is applicable since time intervals with different beginning dates may yield different longitudinal weights. BW_1 applies for the restricted universe, since the longitudinal weight is the cross-sectional household weight as of the first month of the time interval for all households in sample that month, and zero for all other households. However, BW_2 applies for the unrestricted universe since longitudinal households that entered sample after the beginning of the time interval are treated as in the BH procedure.

BD_1 holds since any household with a positive weight was either in sample the first month of the time interval or the month that the household began, and consequently, no retrospective data are needed.

As in the BH procedure, and for the same reasons, FW_1 applies for the unrestricted universe, FW_2 for the restricted universe and FD_2 for both universes.

SH, RM, and SE Definitions, CM Procedure, Restricted Universe. T_2 is applicable since any two intervals may yield different longitudinal weights.

Furthermore, BW_1, BD_1, FW_1, and FD_1 apply. The explanation is similar to that given for the NC definition except now the two cases are:
(a) No original sample people were household members for the entire time interval. (b) At least one original sample person was a household member for the entire time interval.

SH, RM, and SE Definitions, AW Procedure. T_2 is applicable since any two time intervals may yield different longitudinal weights.

Any household that contained an original sample person for at least one month of the time interval receives a positive longitudinal weight for the unrestricted universe, while for the restricted universe it receives a positive weight if it also existed for the entire time interval. However, for either universe such a household might have existed for months when there were no original sample persons in the household, both before and after it came into sample. Hence BD_2 and FD_2 apply. Furthermore, in order to compute the longitudinal household weight it is necessary to determine if the household was in existence at the beginning and the end of the time interval for both universes, and in addition for the unrestricted universe, the beginning and ending dates if they are within the time interval. Hence BW_3 and FW_2 hold.

Utilization of Data. Having compared the procedures with respect to needs for unavailable data and the multiplicity of weights, we now turn our attention to variance. To compare the variance characteristics of the procedures we will focus on the amount of collected data that is used in obtaining estimates, since this is a primary determinant of variance. This discussion will also better illustrate the proportion of data needed for estimation that is unavailable for each procedure. In general, the greater this proportion is, the larger the burden is on any missing data procedure employed, with a resulting greater potential for bias problems. To make the comparison we show in Table 2, all 24 possible cases of how the data on a longitudinal household may be complete, partly available, or nonexistent for a particular time interval.

The symbols t_B, t_1, t_2, and t_E denote beginning date of household, first sample month, last sample month, and ending date of household respectively. The columns indicate different time intervals. Interval B is the interval of interest. Interval A is from t_B until the beginning of interval B, while interval C is from the end of interval B until t_E. The fifth case, for

example, is of a household that formed before interval B about which we are missing some data pertinent to the early part of interval B. The first nine cases comprise the restricted universe. The last 15 cases fill out the unrestricted universe. Each case is marked as having complete data, partial data, or no data. Of course, all of this is assuming perfect response. The only type of missingness that we are discussing here is that caused by operational procedures. On the right there is a column for each procedure with an "A" entered if it always uses the case, an "S" if it sometimes uses the case but not always (which will be explained in the discussion that follows), and a blank otherwise. These comparisons do not apply to the NC definition, for which all five procedures use all the complete cases and no other cases.

Table 2.

Data Utilization

	Interval A	Interval B	Interval C	Completeness	Procedure BH	BI	CM	AW	HW
1	$t_B=t_1$		$t_2<t_E$	perfect	A	A	S	A	S
2	$t_B<t_1$		$t_2<t_E$	perfect		A	S	A	
3	$t_B=t_1$	t_2	t_E	some missing	A	A		A	
4	$t_B<t_1$	t_2	t_E	some missing		A		A	
5	t_B	t_1	$t_2<t_E$	some missing				A	
6	t_B	t_1 t_2	t_E	some missing				A	
7	$t_B=t_1$ t_2		t_E	all missing	A				
8	$t_B<t_1$ t_2		t_E	all missing					
9	t_B		t_1 $t_2<t_E$	all missing					
10	$t_B=t_1$	$t_2=t_E$		perfect	A	A		A	S
11	$t_B<t_1$	$t_2=t_E$		perfect		A		A	
12		$t_B=t_1$	$t_2<t_E$	perfect	A	A		A	S
13		$t_B=t_1$ $t_2=t_E$		perfect	A	A		A	S
14	$t_B=t_1$	$t_2<t_E$		some missing	A	A		A	
15	$t_B<t_1$	$t_2<t_E$		some missing		A		A	
16		$t_B=t_1$ $t_2<t_E$		some missing	A	A		A	
17		$t_B=t_1$ t_2	t_E	some missing	A	A		A	
18	t_B	t_1 $t_2<t_E$		some missing				A	
19		$t_B<t_1$ $t_2<t_E$		some missing				A	
20		$t_B<t_1$ t_2	t_E	some missing				A	
21		$t_B<t_1$	$t_2<t_E$	some missing				A	
22	$t_B=t_1$ t_2	t_E		all missing	A				
23	$t_B<t_1$ t_2	t_E		all missing					
24		t_B	t_1 $t_2<t_E$	all missing					

The BH procedure uses the complete cases 1, 10, 12, and 13, but does not use the complete cases 2 and 11. It also uses the partial cases 3, 14, 16, and 17, and cases 7 and 22 for which there is no data in interval B. The BI procedure uses all the complete cases, more of the partial cases and none of the cases with no data. We thus think the BI procedure will tend to produce smaller variances than the BH procedure since it uses more of the available data. However, it is not clear in general which of these two procedures has the smaller proportion of needed data that is missing.

The CM procedure is appealing for the restricted universe since it uses all the complete cases (except in the rare situation when there is at least one original sample person present for every month of interval B, but none of them are present for the entire interval), and none of the other cases. It should thus have fairly small variances and has only the slight bias indicated in Section 3. However, it is not applicable to the unrestricted universe.

The HW procedure uses the same complete cases as the BH procedure, except it does not use these cases when the householder is not an original sample person, and it uses none of the other cases. However, it is not applicable to the RM, SE, and most other longitudinal household definitions.

The AW procedure is the most aggressive in utilizing partial data. It uses all the complete and partial cases while avoiding the cases with no data. Also note that it assigns smaller weights, in general, to the partial cases than the complete cases. We believe it will tend to produce the smallest variances for most definitions, particularly in the unrestricted universe, but also tends to have the highest proportion of data that is needed for estimation but unavailable.

7. ADJUSTMENTS OF ESTIMATES

In this section we will present some general ideas on adjustments to be made to the unbiased longitudinal household weights that would be obtained using any of the procedures described in Section 3. These should be considered only as preliminary thoughts, as many details remain to be worked out, and even the general approach is subject to change. The proposed procedures are somewhat analogous to the procedures used for cross-sectional estimates, and contain the following four components: an adjustment for the purpose of reducing between PSU sampling variability; an adjustment for household non-interview in second and subsequent waves; and a final adjustment to CPS estimates of the number of households by age-race-sex category of householder.

The first suggested step in the process of adjusting the unbiased weights does not actually begin with these weights, but instead alters the output of Section 3, so the resulting weights contain adjustments for first wave noninterview, and to reduce between PSU sampling variability. To do this, we simply alter the description in Section 3 of the first wave cross-sectional weight to now include these two adjustment factors in addition to the reciprocal of the probability of selection.

Two further adjustments would be performed on the weights resulting from the modification described in the previous paragraph. The need for the first adjustment would arise because there would be longitudinal households resulting from wave one respondent households for which there were missing data, not "completed" by imputation, for at least part of the time interval for which estimates are desired. This adjustment would redistribute the weights of such households to all households in the same weighting cells with complete data, in proportion to the weights of the households with complete data. In performing this adjustment it should be noted that in the case of

households for which complete contact is lost after some point, subsequent household compositional changes may alter the weights of the noninterview households, so it is not always clear what are the correct weights to redistribute. Imputation of these weights would appear to be necessary.

The final proposed adjustment would adjust the SIPP sample estimate of number of longitudinal households whose householder is in a given age-race-sex category to the CPS estimate. This would be accomplished by multiplying each household weight in the given cell by the ratio of the CPS estimate of the number of households in the cell to the SIPP estimate. (Family estimates could be controlled to CPS estimates by further dividing each cell into family and non-family household subcells. Even finer subdivision is also possible.) There are several possible approaches to computing this adjustment factor for each cell. The simplest would be to compute the factors at one month during the time interval in question, where the denominator of the ratio would be the sum of the weights of all longitudinal households in the cell in existence during that month, and then applying that same factor also to all other longitudinal households in the cell. (This was done in NMCUES (Whitmore, Cox, and Folsom 1982).) If this approach is taken then, in general, the SIPP and CPS estimates of the number of households in a given cell, and even the estimated total number of households in the universe, would not agree for any other month.

If it is required that the SIPP longitudinal household estimates in each cell agree with CPS estimates for every month in a time interval, then this could be accomplished by grouping the longitudinal households in each cell according to their pair of beginning and ending dates, and applying a different weighting factor for each such group. The values for these factors could be determined by considering them as variables in a mathematical

programming problem. This is described in detail by Judkins et al. (1984). Caution should be taken before adopting such a technique to control household weights for every month in a time interval. In certain situations no solution would be possible unless some weighting factors were allowed to be very large, or even negative. It may sometimes even occur that there is no solution even when there are no constraints on the weighting factors. Furthermore, slight changes in the objective function or the constraints might dramatically change some weighting factors. Finally, under some of the proposed definitions the householder in a longitudinal household may change, placing the household in a different age-race-sex cell, and requiring a modification of the procedure to account for this problem.

Some necessary imperfections in the CPS household control totals should also be noted. Although the CPS estimates of total individuals in a given age-race-sex category are themselves controlled to independent demographic estimates which have no sampling variability, the number of householders in each category is not controlled in this manner. This is troubling because the process which yields the CPS estimates of households is subject to unknown biases. Despite this, it is felt that this use of CPS estimates in adjusting SIPP data would reduce total sampling variability and many biases because of the combination of the demographic estimate controls and the larger size of the CPS sample.

REFERENCES

Dicker, Marvin and Casday, Robert J. (1982), "A Reciprocal Rule Model for a Longitudinal Family Unit," American Statistical Association - Proceedings of the Social Statistics Association.

Judkins, D.R., Hubble, D.L., Dorsch, J.A., McMillen, D.B., and Ernst, L.R. (1984), "Weighting of Persons for SIPP Longitudinal Tabulations," American Statistical Association - Proceedings of the Section on Survey Research Methods, to appear.

Kasprzyk, Daniel and Kalton, Graham (1983), "Longitudinal Weighting in the ISDP," Technical Conceptual and Administrative Lessons of the ISDP, Social Science Research Council, New York.

McMillen, David B. and Herriot, Roger A. (1984), "Toward a Longitudinal Definition of Households," SIPP Working Paper Series, No. 402, Bureau of the Census, Washington, D.C.

Nelson, D.D., McMillen, D.B., and Kasprzyk, D. (1984),"An Overview of the Survey of Income and Program Participation", SIPP Working Paper Series, No. 401, Bureau of the Census, Washington, D.C.

Whitmore, R.W., Cox, B.G., and Folsom R.E. (1982), "Family Unit Weighting Methodology for the National Household Survey Component of the National Medical Care Utilization and Expenditure Survey," Research Triangle Institute report.

EARLY INDICATIONS OF ITEM NONRESPONSE ON THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

John F. Coder and Angela M. Feldman
U.S. Bureau of the Census

For Presentation at the
American Statistical Association
Annual Meetings - August 1984

EARLY INDICATIONS OF ITEM NONRESPONSE ON THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

by John F. Coder and Angela M. Feldman

Introduction

The Survey of Income and Program Participation (SIPP) promises to become the most important source of data for measuring the level of and changes in the economic well-being of the U.S. population. Collection of these data began in the fall of 1983. The survey design for the initial sample of 25,900 housing units in the noninstitutional population, calls for each household to be interviewed at 4-month intervals over a 2-1/2 year period. The sample is divided into 4 rotations or panels of equal size and one panel is interviewed in each month throughout this period resulting in a total of eight personal contacts by Census interviewers for each sample household.

The first interviews in this new survey were conducted during October, November, and December of 1983, and January 1984. The questionnaire used to collect information in the initial interview concentrates on labor force participation and sources and amounts of income. Most data is recorded separately by month for the 4-month reference period ending in the month prior to the month of interview. For example, data collected in the October 1983 interviews covered the June through September period. Most interviews were completed during the first 2 weeks of the interview month.

The primary purpose of this paper is to present some preliminary indications of the item nonresponse rates for the first interviews of SIPP. These rates of nonresponse cover labor force, income recipiency, and income amounts. The effect of self or proxy respondents on nonresponse rates is discussed for

a selected group of items. Some data on other aspects of the survey have also been included. These are overall household noninterview rates, average times required for interviews, and use of callback procedures to obtain missing information.

Item Nonresponse

Item nonresponse is defined in this paper to mean a missing answer to a specific question that should have been answered. Item nonresponse can result for many reasons, the most frequent being lack of knowledge by the respondent, i.e., "Don't Knows," and refusals to answer. Nonresponse can also result when the interviewer fails to record a response in the correct location or follows an incorrect path within the questionnaire design.

Labor Force Items--Table 1 shows preliminary nonresponse rates for items 2a, 2b, 4, 5a, 5b, 6a, 6b, 6c, 7a, 7b, and 8a of the labor force and recipiency section on the first interview questionnaire. The questions themselves are shown in Figure 1.

In general, the nonresponse rates for the labor force questions were low (see table 1). The nonresponse rate on item 2a, incidence of looking for work or on layoff for persons who did not work at all during the reference period (nonworkers) was only 0.4 percent. About 6.7 percent of the nonworkers reporting looking or on layoff had a nonresponse for item 2b, the number of weeks spent looking or on layoff. The comparable nonresponse rates for workers were 1.0 percent for incidence of looking or on layoff (item 7a) and 3.2 percent for item 7b, the number of weeks spent looking or on layoff. The nonresponse rate for item 4, asking if the respondent held a job or business during the entire 4-month reference period, was less than 0.1 percent.

One of the questions with a relatively high nonresponse rate in the labor force section was item 5b covering the number of weeks absent without pay for persons having a job for the entire period. The nonresponse rate for this question was 11.6 percent.

Item 8a is the question covering the number of hours usually worked per week during the 4-month period. This critical data item was missing for 1.3 percent of the 25,510 sample persons reporting a job or business during the reference period.

Income Recipiency.--A major portion of the questionnaire was designed to determine the sources of income received during the 4-month period by each household member age 15 years old and over. A total of 52 different income sources (other than earnings from employment) were covered in the survey. Tables 2 and 3 show income recipiency nonresponse rates and ratios of nonresponses to "YES" responses for SIPP and the March 1983 CPS for a selected group of income types. The rates refer to the 4-month reference period for SIPP and calendar year 1982 for the March CPS.

The nonresponse rates for SIPP are extremely low and vary only slightly by rotation. The nonresponse rate on recipiency for SIPP ranged from less than 0.1 for Aid to Families with Dependent Children and private pensions to 1.3 percent for stocks or mutual funds. In contrast, the rates for the March 1983 CPS clustered around the 10-percent level. These rates for the March CPS are largely attributable to the 7 percent household noninterview rate on the income supplement questionnaire.

The last two columns of table 3 show the ratios of nonresponses to "YES" responses for SIPP and the March CPS. This measure of nonresponse may be better

than the overall nonresponse rate because it provides a measure that is relative to the size of the recipient universe. The March CPS ratios are again much higher than those encountered in the first interview of SIPP. This difference is also related to the 7 percent March supplement noninterview rate. Given this fixed nonresponse rate the ratio is inversely related to the proportion of the population receiving a specific income type. This is evident by the large ratio of 4.01 for Aid to Families with Dependent Children. The ratio itself means that, in this case, the number of nonresponses and, therefore, imputations required exceeded the number of "YES's" by a factor of 4 to 1.

Hourly Wage Rates.--The nonresponse rates on hourly wages are shown in table 4. These rates are shown separately by type of respondent. The nonresponse rate was 9.5 percent overall, 5.1 percent for self response and 16.7 percent for proxy response. The overall nonresponse rate for hourly wages increased from the 7.8 percent level in October to 10.5 percent in January. This resulted mainly from an increase in the nonresponse rate for proxy responses of from 13.8 percent in October to 19.2 percent in January. Approximately 62 percent of the respondents were "self."

Monthly Wage or Salary Income.--Table 5 contains the nonresponse rates for the monthly amounts of wage and salary income. The nonresponse rate overall averaged about 6.2 percent for the initial SIPP interviews. The rate for self respondents, which accounted for 64 percent of the total, was lower, 4.6 percent, while the rate for proxy respondents was 9.0 percent. The 9.0-percent nonresponse rate for proxy interviews on monthly earnings amounts was considerably lower than the comparable rate of 16.7 percent for hourly wage amounts.

Nonresponse rates increased from 5.4 percent to 6.7 percent between October and January.

Self-Employment Income.--Nonresponse rates for self-employment income have traditionally exceeded those for most income types. The items in the self-employment section of the SIPP questionnaire cover monthly amounts of "salary" and other income received by owners of businesses, professional practices, farms, etc. The question is not designed to obtain estimates of the business's net profit on a monthly accounting period. An additional question was included covering estimated net profit for the entire 4-month reference period. The nonresponse rate overall for the monthly salary or other income received by the self-employed was 14.0 percent (see table 6). The nonresponse rate for proxy interviews exceeded that of self-responses by a considerable margin. The rate for proxy interviews was 22.3 percent compared to 9.8 percent for self responses. Nonresponse rates were slightly higher in January than October, increasing from 13.6 percent to 15.1 percent. About two-thirds of respondents for this item were "self."

Interest Income.--Table 7 contains nonresponse rates for interest amounts received during the SIPP 4-month reference period. These rates cover the interest amount received from one or more of the following sources: 1) regular or passbook savings, 2) money market deposit accounts, 3) certificates of deposit, or other savings certificates, and 4) NOW accounts or other interest earning checking accounts. The nonresponse rate for interest income from these sources was 34.6 percent. The rate in January was 35.4 percent, somewhat higher than the 32.6 percent for October. About 4 percent of the total number of nonresponses on interest amounts can be attributed to refusals. The remainder

were mainly categorized as "Don't Knows." A "Don't Know" response to interest income was followed by a question to obtain the balance or amount in the account. The nonresponse rates for this item are also shown in table 7. The nonresponse rate for balances in savings was 24.2 percent. In combination these two nonresponse rates indicate that both the interest amount and the balance amount were missing in only about 13.3 percent of the sample cases for these sources of interest income.

Dividend Income.--The questions covering the amount of dividend income received were divided into two categories, those dividends actually received and those credited against a margin account or automatically reinvested in additional shares of stock. As indicated by the data in table 8, the nonresponse rates for these two categories differ significantly. The rate for dividends actually received was 9.4 percent. The rate for dividends credited was 30.7 percent.

Noninterview Rates

The noninterview rate is a measure of the proportion of occupied housing units, i.e., those eligible for interview, for which interviews were not obtained. As mentioned earlier the total sample size for the 1983 SIPP was about 25,900 housing units. Of this total about 4,600 were not eligible for interview. These ineligible units were found to be vacant, demolished, under construction, or unoccupied for other reasons. This left 19,900 households eligible to be contacted. Interviews were not obtained for 4.8 percent of this group (see table 9). Most noninterviews, about 77 percent, were refusals to participate. The remainder of the total noninterview rate consisted of

situations classified as "no one home" and "temporarily absent." These classifications were assigned after repeated visits failed to yield a contact.

The noninterview rate varied considerably by region of the Country. The lowest noninterview rate was 2.4 percent from the Kansas City Regional Office that covers Kansas, Missouri, Iowa, Minnesota, and Wisconsin. The highest noninterview rate was 10.1 percent from the New York Regional Office covering the parts of New York and New Jersey in the vicinity of New York City.

There was slight variation in the noninterview rates by month of interview, however, there was no apparent trend. The rate for the first month of interview was 5.1 percent compared to 4.3 percent, 5.2 percent, and 4.8 percent in the succeeding 3 months, respectively. The overall noninterview rate of 4.8 percent was somewhat higher than the March 1983 CPS rate of 4.4 percent. The rate for SIPP was, however, lower than the 5.4 percent noninterview rate for the panel coming into the March 1983 CPS for the first time. As noted earlier, about 7.0 percent of the March CPS sample households completed the monthly labor force questions but were noninterviews on the income supplement. These cases are in addition to the 4.4 percent household noninterviews.

Callback Items

The design of the SIPP questionnaire incorporated procedures for following up on missing responses to items identified as either especially important to the overall quality of the survey data or with previously noted high nonresponse rates. The first step in this process was the determination that the answer to the designated question would be available from another household member not present at the time of the interview or at a later date. If so, the interviewers, in most cases, called back by telephone to obtain the missing information. The data in table 10 summarize use of the callback system.

The callback system appears to be most effective for obtaining missing data on amounts of monthly wage and salary income. About 600 cases were marked for callback for these amounts. The procedure obtained responses to the missing earnings amounts in about 7-out-of-10 cases.

Use of the callback was less successful in obtaining missing amounts for the other income sources. Slightly more than half (54 percent) of the callbacks were successful for obtaining data for the monthly amount of salary and other income received from self-employment. Attempts to follow up on amounts of interest and dividend income from various sources proved to be even less effective. About 45 percent of the respondents were able to supply an amount when contacted by an interviewer. Use of the callback procedures appears to have declined between the October and January interviews. The number of cases marked for follow-up in January were significantly lower than October for each income type. While less frequent use of the callback might have been related to a reduced need for follow-up, nonresponse rates for these income types tended to increase between October and Janaury, indicating the opposite.

Interview Time

The time required to conduct an initial SIPP interview is potentially quite long given the number of questions. Obviously households with a large number of adult members, those 15 years old and over, are those that are exposed to the longest overall interview times, on average. The data in table 11 provide the first estimates of interview times based directly on times entered on each person's questionniare by the interviewers. The time required to complete the household control card and roster was added to the interview time on the first questionnaire for the household. These estimates are shown by size of household for the first interview period of SIPP.

The median interview time was 43 minutes for all households in the first interview. The median interview time declined steadily from 48 minutes in October to 41 minutes in January. The median household interview time for 1-person households was about one-half hour while that for 4-person households was one hour and ten minutes. Households with 5, 6, and 7 or more members required proportionally more time for interviews.

Summary

This examination of some of the early "returns" from the 1983 SIPP are, for the most part, encouraging. The household noninterview rate was lower than most had anticipated. The item nonresponse rates were much lower than those experienced in the March CPS. Proxy responses caused significantly higher nonresponse rates for some of the key items studied.

There is reason for concern, however, in several areas and these should be watched closely. The first is the general trend toward higher nonresponse rates between October and January interviews. The second is the relatively high noninterview rate for the New York area. While this is consistent with our experiences in other surveys, this rate should be monitored closely as will the rates in the other regions.

The next step in the evaluation of the 1983 SIPP data will be comparison of the survey estimates of income recipients with figures derived from program statistics and other independent sources. This analysis will provide a very important look at the magnitude of survey underreporting, a major concern of SIPP and other household income surveys.

Figure 1. Selected Labor Force Questions

NONWORKERS

2a. Even though ... did not have a job during this period, did ... spend any time looking for work or on layoff from a job?

|__| YES -- ASK 2b

|__| NO

2b. In which weeks was ... looking for work or on layoff from a job?

WORKERS

4. Did ... have a job or business, either full or part time, during EACH of the weeks in this period?

|__| YES -- ASK 5a

|__| NO -- ASK 6a

5a. Was ... absent without pay from ...'s job or business for any FULL weeks during the 4-month period?

|__| YES -- ASK 5b

|__| NO

5b. In which weeks was ... absent without pay?

WORKERS WITH WEEKS WITHOUT A JOB OR BUSINESS

6a. In which weeks did ... have a job or business?

6b. Was ... absent from work for any full weeks without pay?

|__| YES --ASK 6c

|__| NO

6c. In which weeks was ... absent without pay?

7a. During the weeks that ... did not have a job did ... spend any time looking for work or on layoff?

|__| YES -- ASK 7b

|__| NO

7b. In which of these weeks was ... looking for work or on layoff from a job?

WORKERS

8a. In the weeks that ... worked during the 4-month period, how many hours did ... usually work per week?

Table 1. Selected Item Nonresponse Rates for the Labor Force Items on the 1983 SIPP: Interview No. 1

Item	Total	Rotation			
		One	Two	Three	Four
2a	0.4	0.4	0.4	0.4	0.3
2b	6.7	8.2	6.8	5.9	5.9
4	0.1	0.1	0.1	(Z)	0.1
5a	0.1	0.1	0.1	0.1	0.1
5b	11.6	12.6	11.0	8.2	14.4
6a	2.2	2.9	2.0	1.9	1.8
6b	3.3	6.6	2.3	1.8	1.4
6c	6.8	2.1	12.2	3.3	10.5
7a	1.0	0.9	1.0	1.1	0.9
7b	3.2	4.7	3.7	2.0	2.0
8a	1.3	1.3	1.3	1.2	1.2

Z Less than .05 percent.

Table 2. Selected Item Nonresponse Rates for Income Recipiency During the 4-month Reference Period on the 1983 SIPP: Interview No. 1

Income type	Total	Rotation			
		One	Two	Three	Four
Social Security.....	0.6	0.6	0.6	0.5	0.5
Unemployment compensation..	0.1	0.1	0.1	0.1	0.1
Veteran's payments.....	0.2	0.2	0.2	0.2	0.2
Aid to Families with Dependent Children.....	(Z)	(Z)	(Z)	(Z)	(Z)
Food stamps.....	0.3	0.4	0.4	0.2	0.2
Private pensions.....	(Z)	(Z)	(Z)	0.1	(Z)
Savings accounts.....	1.0	0.8	0.8	1.1	1.1
Shares of stock or mutual funds.....	1.3	1.4	1.1	1.3	1.3
Rental property.....	1.0	1.0	0.7	0.9	1.1

Z - Less than .05 percent.

Table 3. Selected Income Nonresponse Rates from the March 1983 CPS, Ratio of Nonresponses to "YES" Responses for the March 1983 CPS, and Ratio of Nonresponses to "YES" Responses for Interview NO. 1 of the 1983 SIPP

Income type	March 1983 CPS nonresponse rate	March 1983 CPS ratio of nonresponses to "YES's"	1983 SIPP ratio of nonresponses to "YES's"
Social Security.....	9.6	0.61	.03
Unemployment compensation..	9.6	1.16	.03
Veteran's payments.....	9.6	1.14	.10
Aid to Families with Dependent Children.....	9.7	4.28	.01
Food stamps.....	6.4	0.84	.07
Private pensions.....	9.6	1.64	.01
Savings accounts.....	10.4	.215	.02
Shares of stock or mutual funds.....	9.7	0.69	.09
Rental property.....	9.7	0.66	.13

Table 4. Nonresponse Rates on Hourly Wage Rate by Type of Respondent for the 1983 SIPP: Interview No. 1

Type of respondent	Total	Rotation			
		One	Two	Three	Four
Total.....	9.5	7.8	9.3	10.4	10.5
Self.....	5.1	4.1	4.7	5.9	5.6
Proxy.....	16.7	13.8	16.1	18.0	19.2
Proportion of Self Responses.	.62	.62	.60	.63	.64

Table 5. Nonresponse Rates on Monthly Wage and Salary Income by Type of Respondent for the 1983 SIPP: Interview No. 1

Type of respondent	Total	Rotation			
		One	Two	Three	Four
Total.....	6.2	5.4	5.8	6.8	6.7
Self.....	4.6	4.2	4.3	4.9	4.9
Proxy.....	9.0	7.6	8.4	10.2	10.1
Proportion of Self Responses.	.64	.63	.63	.64	.65

Table 6. Nonresponse Rates on Monthly Amounts of Self-Employment Income for the 1983 SIPP: Interview No. 1

Type of respondent	Total	Rotation			
		One	Two	Three	Four
Total.....	14.0	13.6	12.6	14.6	15.1
Self.....	9.8	9.5	9.7	9.6	10.2
Proxy.....	22.3	21.4	18.6	24.3	24.7
Proportion of Self Responses.	.66	.65	.67	.66	.66

Table 7. Nonresponse Rates for Amounts of Interest Income from the 1983 SIPP: Interview No. 1

Item	Total	Rotation			
		One	Two	Three	Four
Interest amount.....	34.6	32.6	33.8	37.1	35.4
Percent refusals.....	4.2	4.1	4.0	4.6	4.1
Balance amount.....	24.2	23.6	24.1	24.9	24.1

Table 8. Nonresponse Rates for Amounts of Dividend Income for the 1983 SIPP: Interview No. 1

Item	Total	Rotation			
		One	Two	Three	Four
Dividends received.....	9.4	10.3	8.3	9.8	9.3
Dividends credited.....	30.7	28.2	33.8	30.1	30.5

Table 9. Household Noninterview Rates by Regional Office for the 1983 SIPP: Interview No.1

Item	Total	Rotation			
		One	Two	Three	Four
Total.....	4.8	5.1	4.3	5.2	4.8
Boston.....	3.8	2.9	2.5	5.4	4.6
New York.....	10.1	13.3	8.3	10.8	8.4
Philadelphia.....	3.0	2.0	3.4	2.5	4.1
Detroit.....	4.1	3.0	3.6	5.4	4.1
Chicago.....	4.8	5.0	3.4	5.7	5.0
Kansas City.....	2.4	1.6	1.6	4.0	2.5
Seattle.....	4.7	5.1	4.4	5.2	4.3
Charlotte.....	3.5	4.3	2.7	2.8	3.8
Atlanta.....	4.9	5.4	5.0	5.2	4.2
Dallas.....	5.1	5.0	5.1	4.6	5.8
Denver.....	5.3	6.1	5.7	4.1	5.5
Los Angeles.....	7.5	9.3	6.2	8.9	5.8

Table 10. Success Rates of Callback Items

Item	Total	Rotation			
		One	Two	Three	Four
Success Rates					
Wages and salary.....	71.0	76.2	76.9	70.0	59.0
Self-employment.....	54.0	58.6	55.0	48.3	54.5
Interest and dividends.	44.8	48.4	49.6	38.2	40.8
Number of Callbacks					
Wages and salary.....	599	172	143	150	134
Self-employment.....	100	29	20	29	22
Interest and dividends.	582	192	139	131	120

Table 11. Median Household Interview Times by Number of Members 15 Years Old and Over from the 1983 SIPP: Interiew No. 1

Number of persons	Total	Rotation			
		One	Two	Three	Four
Total.....	43	48	44	42	41
One.....	29	33	30	26	26
Two.....	44	50	45	42	41
Three.....	57	64	57	55	55
Four.....	70	76	72	67	66
Five.....	83	90	81	84	77
Six.....	98	105	111	101	71
Seven or more.....	113	114	(B)	120	94

B Less than 10 sample households.