HOUSEHOLD INCOME AND WEALTH: SAMPLE DESIGN AND ESTIMATION FOR THE 1983 SURVEY OF CONSUMER FINANCES

Steven G. Heeringa and Richard T. Curtin, University of Michigan

I. INTRODUCTION

Sample surveys are an important source of data for the study of household financial characteristics, but they present the researcher with a number of difficult design and estimation problems. For univariate analyses the root of the problem is the asymmetric, highly skewed population distribution of income and asset variables. Multivariate analyses of survey data are influenced both by the distributional properties of individual variables and the weak and sometimes highly irregular relationships among various forms of household income and financial and non-financial assets. This paper draws on data collected in the 1983 Survey of Consumer Finances (SCF) to illustrate the sampling and estimation problems that are common to income and wealth surveys and to review methods designed to address these problems.

The focus of this paper is primarily on the body of the income or wealth distribution. The intent here is not to provide in-depth treatment of estimation for the open ended category of those very high income or wealthy households which occupy the tip of these distributions' upper tails. Although no attempt is made to draw an exact boundary between the wealthy and the not so wealthy, design and estimation problems encountered in the extreme upper ranges of income and wealth distributions require special and possibly model-based solutions which are the subject of separate papers at this conference.

Including these introductory remarks, this paper is organized into seven sections. To highlight important sampling and estimation problems for income surveys, Section II describes a statistically optimal solution to the sample design problem and contrasts the properties of such a design with those of practical alternatives. Taking as an example the 1983 Survey of Consumer Finances, Section III continues the discussion of practical dual frame sample designs for studies of household income and wealth. Current weighting and estimation alternatives for the 1983 SCF are covered in Sections IV and V. Section VI deals with sampling variance properties of 1983 SCF estimates of income statistics. A summary is presented in Section VII.

II. STATISTICAL FRAMEWORK FOR THE SURVEY DESIGN. THEORETICAL OPTIMA VS. PRACTICE.

The purpose of this section is to briefly outline a "textbook" or theoretically optimum approach to the sampling and design-based[1] estimation of household income and wealth characteristics and to contrast the features of a desired optimum with those of operational sample designs which must conform to a variety of practical constraints.

II.A. Stratified Sampling With Optimal Allocation

Based on the theory of stratified sampling, an optimal design would begin by clearly identifying each element of the survey population (frame identification). Based on known characteristics of individuals, highly correlated with the variables of greatest interest to the study (income, net worth), strata of elements would be formed and each sample element would be uniquely assigned to a stratum (stratification). If the objectives of the study could be refined to interest in a single continuous variable (or possibly two), optimal stratification (Dalenius, 1957) or the stratification based on the cumulative square root of f_y rule (Cochran,1963) could be used to define stratum boundaries.

The optimal design approach requires that the stratum population totals, N_h (h=1,...H) are known and that a unique stratum identification can be assigned to each population element in the sample frame. Based on the stratum population size and element variance for the characteristic of interest (or a highly correlated stratifier), sampling rates for individual strata would be set according to the standard Neyman allocation formula (allocation):

$$f_h = \frac{n_h}{N_h} = KW_h S_h \text{ ; where } W_h = N_h/N.$$

If the distribution of the characteristic of interest is highly skewed at its upper tail, the "optimal" sampling fraction for the strata of highest values would in all likelihood be equal to or greater than 1. For these strata, all elements will be included in the sample with certainty— the very rich would enter the sample with probability 1. Elements in other strata would be randomly sampled with probability equal to the sampling fraction, f_h, determined under the optimal allocation.

Under the optimal design plan, unbiased estimates of the mean per element statistic and its estimated sample variation are computed using the stratified estimators:

$$\bar{y}_{st} = \sum_{1}^{H} W_h \bar{y}_h \text{ ; and}$$

$$\text{var}(\bar{y})_{st} = \sum_{1}^{H} (1-f_h) W_h^2 S_h^2 / n_h .$$

II.B. Constraints on Practical Survey Design

II.B.1. Sample frames.—Step one in the construction of the optimal design is the complete definition of the population of elements — a perfect frame is presumed available for the probability selection of elements. Ideally, the perfect sampling frame would be a single list with an accurate and unduplicated entry for each population element. For a given tax year, a complete list of federal tax filers might be viewed as just such a frame. Unfortunately, access to IRS data bases is highly restricted by confidentiality provisions of U.S. tax law. In the 1983 SCF, the Internal Revenue Service (IRS) was barred from providing a list frame of tax filers directly to the Survey Research Center (SRC). Instead, IRS selected a sample of tax filers from its 1980 Statistics of Income data base and only after obtaining written signed consent, released the names and addresses of cooperating sample taxpayers to SRC.

Even with IRS assistance in gaining controlled research use of the SOI tax files, problems of timeliness and unit definition still remained. Realistically, computerized data bases of tax filer information would not be available until almost two years after the close of the tax year for which a return is filed. During the intervening period, the taxpayer population would undergo significant change due to deaths, marriage, divorce, influx of new earners, etc. In addition to original noncoverage and increasing obsolescence of the tax filer list over time, the definition of the survey's observational unit (e.g. households, individuals) may differ from the tax filer units which comprise the listed population. For the most part, there should be a good correspondence between tax filings and household income units (particularly in the medium and upper income brackets), but there are a significant number of exceptions and in no circumstance should the two types of units be simply equated.

The basic alternative to the list frame is the use of area probability sampling techniques. In theory, area probability frames will provide complete and unduplicated coverage of households. In practice, area sample coverage of households is

high but less than complete. Errors in listing, definitional problems, transient populations, and inability to access restricted or secured housing areas all contribute to area sampling undercoverage. Despite the advantage of its high degree of population coverage, the area probability frame provides little or no detailed financial information at the level of the individual sampling unit.

A dual frame survey, that is, one which integrates the low information, high coverage properties of area sampling with the high information, unknown coverage properties of tax filer lists, has strong intuitive and statistical appeal. In theory, the dual frame survey provides both proportionate coverage of the population and also provides information needed for optimal (non-proportional) allocation of the sample to population strata. As will be pointed out later, the list frame component of the dual frame design is a particularly valuable tool for disproportionately sampling higher income strata where the variance of financial characteristics reaches its greatest levels.

II.B.2. Stratification variables.—The second major requirement of optimal sampling design for income and wealth surveys is knowledge of the distributional characteristics of the variable of interest or of another variable that is highly correlated with it. In income and wealth surveys, the search for stratification variables focuses on income or income related characteristics of sample units.

For the area probability sample frame, a cost effective income stratification is difficult to achieve. At best, Census data on average household income will enable the sampling statistician to assign area sampling units—tracts, blocks, enumeration districts— to broadly defined income strata. Under the area probability approach, a more refined income based stratification would involve an expensive and procedurally difficult screening of households prior to interview.

Lists frames—specifically Federal tax filer lists—provide detailed income data for stratification; however, even with access to such a high quality source of information on sample elements, previously mentioned problems of timeliness, unit definition and variable definition remain. Tax filer units do not bear a one-to-one correspondence to household units. Source files such as those produced by the Statistics of Income (SOI) program may be two years out of date by the time they could be used as a sampling frame. Form 1040 income definitions and income reporting may differ from that of the survey. In the case of the 1983 SCF, researchers face an added problem arising from the legal restrictions which prevent the IRS from disclosing details about the high income population or the sample selected from the SOI list frame.

II.B.3 Optimal allocation.—A major practical constraint on sample designs for income and wealth characteristics is that the planning of such a design rarely takes place in a univariate statistical setting. Often, data on income and wealth must be collected in a larger multi-purpose survey context. Even in studies such as the 1983 SCF where the primary focus is income, assets, pensions and other financial characteristics of households, pursuit of optimal design characteristics would set up a competition among variables. A sample designed to be optimal for the estimation of household income may not be optimal for the estimation of household net worth. Furthermore, relationships among survey variables may vary from stratum to stratum.

Based on 1983 SCF sample observations, Table 1 presents the estimated correlation between a variable which measures adjusted gross income (AGI) and a selected set of other variables including individual income sources and total net worth. From the table, total sample correlation between AGI and the selected variables is uniformly high; however the strength of these correlations fades as the sample is divided into smaller and smaller income domains. The size of the total sample correlations suggests that multi-purpose stratification of sample elements on the basis of AGI is certainly warranted. The trend toward weaker correlation as the income ranges are restricted indicates that there is little gain in a stratification plan which incorporates many strata based on relatively narrow AGI ranges. In addition to the observed attenuation of correlation as income ranges are narrowed, the pattern of correlation between AGI and other variables changes from one domain to another. The correlation of wage and salary income to total AGI is very high in the <$100K AGI domain and declines steadily across the higher income brackets. In the $500K+ AGI domain, AGI appears to be totally uncorrelated with wage and salary income.

Table 1.--Estimated Correlation between Adjusted Gross Income (AGI) and Major Income Variables and Net Worth*

Variable	Total Sample (n=4103)	Adjusted Gross Income				
		<100K (n=3632)	High Income Categories			
			Total (n=471)	100-199K (n=182)	200-499K (n=190)	500K+ (n=99)
Wages and Salary	.4552	.7221	.2214	.2730	.1930	.0099
Profession, Business	.4546	.3758	.2801	.1575	.0851	.0450
Nontaxable Interest	.4934	.2278	.3970	.0676	.1465	.2791
Taxable Interest	.6123	.3217	.5394	-.0588	.0785	.3882
Dividends	.4884	.2321	.3657	.0940	.2261	-.0059
Sales of Bonds	.6520	.2107	.6653	-.0093	.1029	.6758
Rent and Trusts	.4796	.1704	.4724	.0650	.1330	.3889
NET WORTH	.4997	.1802	.4007	.0908	.1665	.3036

*Correlation estimates (unweighted) from the 1983 SCF.

III. THE 1983 SURVEY OF CONSUMER FINANCES (SCF). A DUAL FRAME SURVEY OF U.S. HOUSEHOLD INCOME, ASSETS AND WEALTH

The 1983 Survey of Consumer Finances, conducted by the Survey Research Center at The University of Michigan, continued a longstanding research program on household income and wealth. The 1983 SCF collected detailed data not only on the amounts and types of financial and nonfinancial assets and liabilities but also on individuals' entitlements to retirement pension benefits.

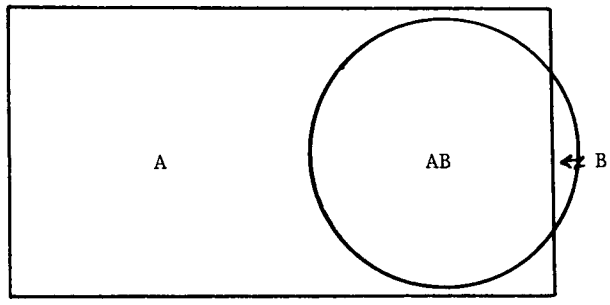
Using a dual frame sampling approach, the 1983 SCF incorporates two overlapping samples of U.S. households and taxpayer units. The first and largest of these is a national area probability sample of U.S. households selected from the Survey Research Center's National Sample Design. Under this multi-stage "cross-sectional" design, each household in the coterminous United States received an equal probability of being selected for interview. The final data set contains n=3665 interviews with area probability sample households. The following discussion will label the national area probability "cross-section" sample in abbreviated form as the "XS" design.

The second sample for the 1983 SCF is a special supplement of higher income tax filers selected from a recent IRS Statistics of Income (SOI) data set. Due to legally imposed constraints which prohibit the IRS from releasing taxpayer information without prior written consent, the IRS cannot provide a detailed description of the sample design for this list sample of high income taxpayers. For the current presentation, it is sufficient to state that the high income sample represents a stratified subselection of SOI sample taxpayers chosen from within a primary stage sample of United States' SMSA's[2] and counties. Stratification of the sample bears an approximate relationship to tax filer income, and the sample allocation is disproportionate across the higher income strata. Additional detail of the stratification and sample allocation plan remains known only to the IRS. In the following discussion, the abbreviation "HY" will be used to reference the high income sample design.

In theory, the area probability sample frame of the XS sample design should provide complete coverage of tax filers represented in the SOI frame. In a highly schematic way, Figure 1 describes the relationship of coverages for the two sample components of the dual frame design. Note that Figure 1 has been deliberately drawn to suggest that in practice the area probability frame may not be perfectly inclusive of all elements in the SOI list frame.

Ignoring for the moment the issue of area sample noncoverage of the high income individuals in the SOI frame (which in theory should be zero), the real ambiguity lies in deciding the exact location of the boundary separating sample units in the XS sample frame which are eligible for the HY sample from those which are not. The uncertainty arises from three sources—one legal, one definitional and one temporal. Under the law, the IRS is bound to protect the confidentiality of tax filer data. To ensure that no illegal disclosure occurs, IRS has not been able to share the exact criteria used to form individual strata of tax filers for the HY sample. There is a general sense that the stratification was related to an AGI measure and that the lower cut-off for HY sample eligibility falls somewhere near $100,000 AGI. Even if IRS could disclose the true nature of the strata used in selecting the HY sample, the definitional and temporal ambiguity would remain. Is the Form 1040 measure of AGI comparable to that obtained in the survey? Strata are based on AGI and related income characteristics reported for the 1980 tax year. Can 1980 tax data be recalled or recovered reliably in the course of the 1983 survey interview? If not, can we assume a stable population distribution for the HY sample strata?

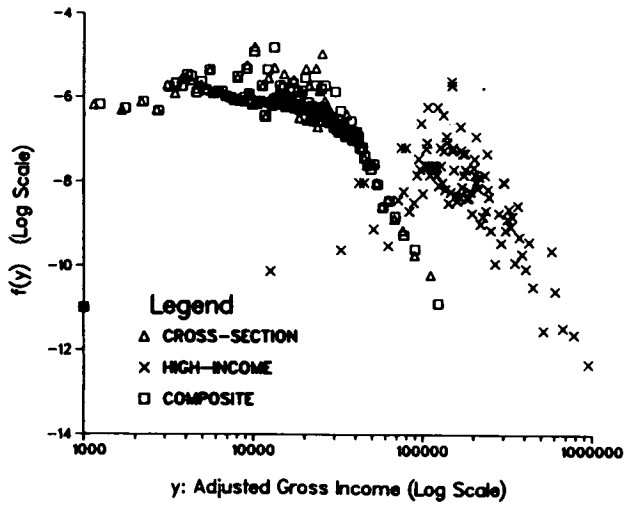
Figure 1.--Schematic Representation of Population Coverage Under the 1983 SCF Dual Frame Sample Design



Where: A = Population covered only by the XS sample frame;
AB = Population covered under both the XS and HY sample frames;
B = Population covered by the HY sample frame but in the zone of noncoverage for the XS sample frame.

Figure 2 is useful for describing the problem which results from uncertainty over the exact "boundaries" of the high income sample strata. The conventional XS sample yields large numbers of sample households in the large body of the income and wealth distribution, but the numbers of observations on higher AGI households will be small. In the 1983 SCF, 64 (1.7%) of 3665 XS sample cases reported 1982 AGI of $100,000 or more. The empirical distribution function of XS sample observations plotted in Figure 2 shows the expected rapid decline in numbers of observations as the $100,000 AGI level is approached. By introducing the supplemental HY sample of n=438 higher income individuals, we increase the count of observations in the upper income range. Note, however, the considerable overlap of the empirical distributions for observations on the two independent samples.

Figure 2.--Empirical Probability Density Functions for Weighted AGI Values from the 1983 Survey of Consumer Finances



IV. WEIGHTED ESTIMATION OF INCOME AND WEALTH

The dual frame sampling and interview of a probability sample of high income tax filers address the common survey problem of a sparsity of observations in the highly critical upper ranges of the income and wealth distributions. The problem now is one of estimation. How are these supplemental data to be used in the estimation of the income and wealth characteristics for the population as a whole?

Originally, SRC ignored the issue of combining the two data sets in analysis, assuming that the XS sample data set would be used for most SCF analyses of households in the body of the income distribution and the HY sample data set would be reserved for independent analysis of the household income and assets of taxpayer units in the upper tail of the income distribution. If a critical need to combine estimates from these two sets of analyses arose, the assumption was that special dual frame estimators or "composite estimators" which reflect the unique error properties of the two independent samples would be used.

Hartley (1962) develops optimal estimators for the means and variances of samples selected under dual frame designs. In very general terms, the dual frame estimators proposed by Hartley involve optimally weighted combinations of independent estimates from the separate frames with appropriate allowance for the "overlap" of the two frames. Most rigorously, the derivation of the optimal weights for the dual frame estimator relies on specific knowledge of: 1) frame boundaries and population counts for each area of coverage (Zones A,B, and AB in Figure 1); and 2) population variances for the characteristics of interest for each zone. Even if the requirements of the dual frame estimator are relaxed through substitution of sample estimates of variances and external estimates of population counts (say from the Statistics of Income program), ambiguity over frame boundaries in the 1983 SCF design is problematic for the application of the dual frame estimator.

Another method of addressing the overlapping coverage of the two frames is to identify their intersection and remove observations in the intersection set from one or the other frame. Here, we might decide to filter out any XS sample cases which were eligible for the HY sample. The barrier to using this approach in the 1983 SCF is that IRS is not able to release the criteria which it used to define the strata boundaries for the HY sample population.[3] Even if disclosure of the stratifying detail were possible, a correct determination of the XS sample cases' eligibility for the HY sample could only be made from their tax return for the appropriate SOI sample year – e.g., 1980 tax year for the SOI subsample which consented to be interviewed in the 1983 SCF.

Soon after the first releases of the 1983 SCF XS and HY sample data sets, researchers expressed a strong interest in developing a single weight value which would permit them to conduct combined or joint[4] analysis of the 1983 SCF XS and HY sample data. With certain strong assumptions, the special dual frame estimators with optimal properties for single estimates could be used to integrate the two overlapping data sources; however, the practicality of these estimators would be diminished by the multi-purpose nature (many variables, many estimates, many statistical procedures) of the survey.

Despite strong reservations due to limited knowledge of the properties (stratification, population sizes, sampling rates, nonresponse) of the HY sample, SRC staff experimented with alternative approaches to develop a single weight variable which analysts could use for joint analysis of the XS and HY sample data sets. To meet the general purpose needs of analysts, this simple weight was constructed by taking the inverse of each individual case's joint probability of being observed under the XS and HY sample designs. We have taken the liberty of using the ill-defined term, "sample observation probability," to refer to an individual's probability of being sampled and, conditional on being sampled, his probability of responding to the survey. Of course, the latter concept assumes the existence of a response probability model that operates within fairly narrowly defined groupings of the sample population (i.e., grouping being the nonresponse adjustment cells used in conjunction with the XS and HY sample data). For lack of a better term, weight values for joint analysis of the 1983 SCF have been labeled the "composite weight" variables. A general description of the original 1983 SCF composite weight variable and a more recent revised version of the original composite weight are given in the subsections which follow.

V. WEIGHTS FOR THE 1983 SCF DATA SET

Original estimation weights for the 1983 SCF XS sample cases were developed by the Survey Research Center. Included in the XS sample weight were factors for: 1) household selection probability; 2) PSU[5] level nonresponse adjustment; and 3) post-stratification to 1980 Census household totals for SMSA/Non-SMSA domains within the four Census regions. Original case weight values for the HY sample were provided to SRC by the Internal Revenue Service. Due to legal complications surrounding the question of what did and did not constitute a possible violation of disclosure regulations, the Internal Revenue Service was prevented from offering real assistance in the development of the original composite weight variables for the joint analysis of the 1983 XS and HY sample data sets.

V.A. Original 1983 SCF Composite Weight Variable.-- Construction of the original composite weight variable required the following set of approximations and assumptions:

1) Each HY sample case was also eligible for XS sample selection. The XS sample observation probability for each HY sample case is equal to the reciprocal of the overall average of XS sample weights for the XS sample data cases.
2) XS sample cases with 1982 AGI of less than $100K were not eligible for the HY sample. In the combined sample, their sample observation probability is proportionate to the inverse of their XS sample weight (i.e., HY sample observation probability is zero).
3) XS sample cases with 1982 AGI greater than $100K were eligible for the HY sample. XS sample observation probabilities for these cases are set equal to the inverse of their known XS sample weight. The unknown HY sample observation probability for these cases is assumed equal to the reciprocal of the modal weight value for HY sample cases reporting 1982 AGI in the same range: $100–199K, $200–499K, $500K+.

Table 2 provides XS and HY sample size counts and average sample-specific weights for respondents categorized into four ranges of reported 1982 AGI. For these same four AGI ranges, Table 3 summarizes the assignment of frame-specific sample observation probabilities to XS sample and HY sample cases.

The joint sample observation probability for each case was computed by adding the assigned XS and HY sample probabilities outlined in Table 3. For example, a cross-section sample case with an XS sample weight of 21,000 and reporting 1982 AGI of $140,000 would be assigned a joint probability of 1/21,000 + 1/6530 = 1/4981. Preliminary values of the composite weight for each sample case were computed by taking the reciprocal of the joint probability sum (e.g., 1/(1/4981)= 4981). As a final control, composite weights were controlled to XS sample based estimates of total households: 1) households with <$100K AGI and 2) households with $100K+ AGI.

What is the general effect of the original composite

Table 2.--Sample Sizes and Average Values of Original Weights for the 1983 SCF XS and HY Samples

AGI Range	XS Sample		HY Sample		
	Sample Cases	Average XS Weight	Sample Cases	Average HY Weight	Modal Value*
$0-$99K	3760	20858	69	3789	---
$100-199K	48	22149	133	4789	6530
$200-499K	14	23857	147	3078	3421
$500K Plus	2	24276	89	1067	310
TOTAL	3824**	20887***	438	3301	---

* Used to assign HY sample observation probability to XS sample cases in the AGI range. (See Table 3.)

** Includes 159 cases which were later deleted because of incompleteness and/or poor quality of the interview data.

***Used to assign XS sample observation probability for HY sample cases in all AGI ranges.

Table 3.--1983 SCF Original Composite Weight Development

Assigned "Sample Observation Probabilities" by Sample Type, AGI Range

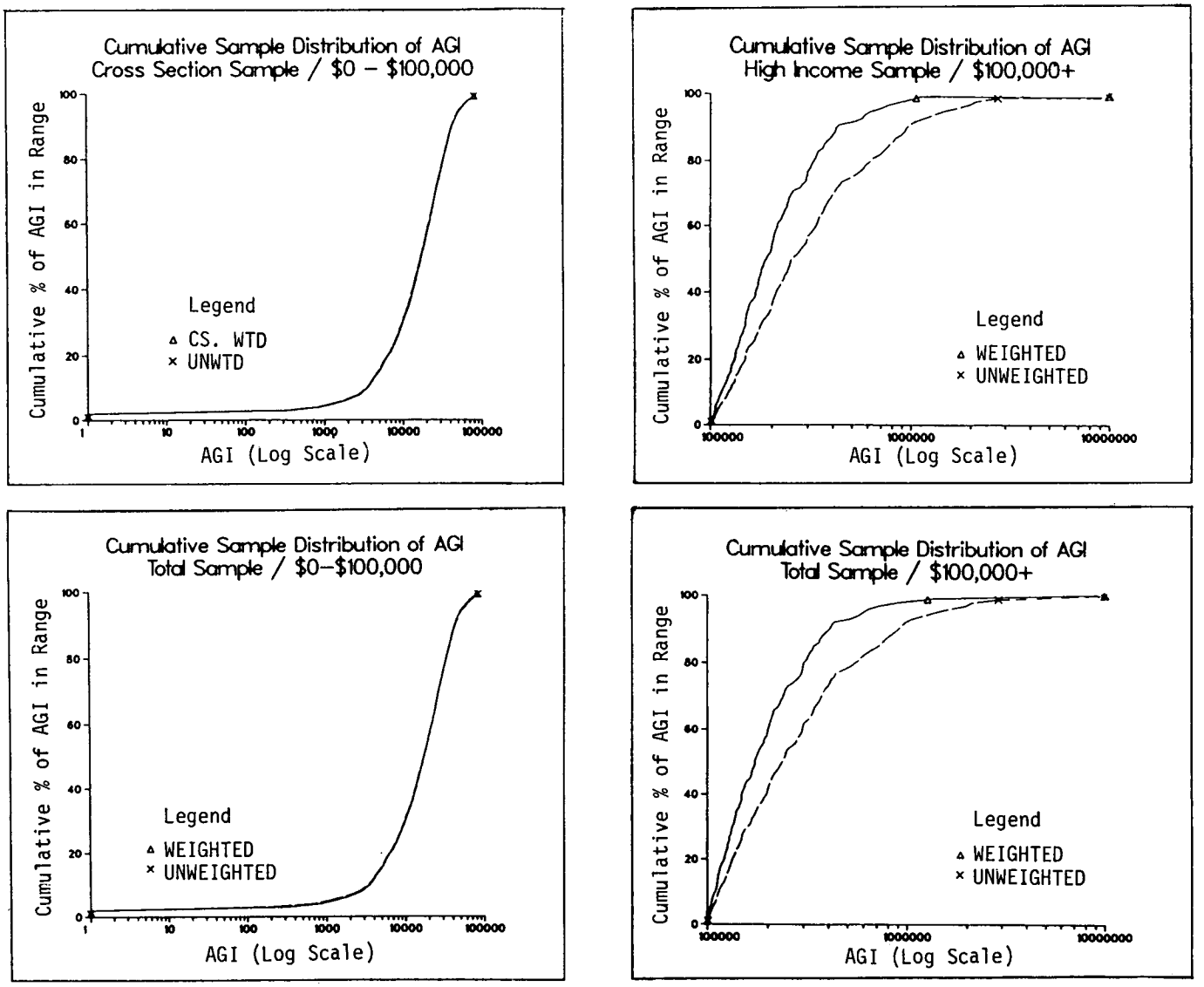
SAMPLE FRAME	PROB UNDER	1982 Adjusted Gross Income Range			
		<$100K	$100-199K	$200-499K	$500K+
XS	XS	1/XS WGT	1/XS WGT	1/XS WGT	1/XS WGT
	HY	0	1/6530	1/3421	1/310
HY	XS	1/20887	1/20887	1/20887	1/20887
	HY	1/HY WGT	1/HY WGT	1/HY WGT	1/HY WGT

weighting on the estimated distribution of AGI in the 1983 SCF data set? Splitting the AGI distribution at $100K, the graphs presented in Figure 3 attempt to answer this question. On the left hand are two graphs which compare the weighted and unweighted distributions of households with 1982 AGI less than $100K. The graph in the upper left-hand corner compares the weighted and unweighted distributions for XS sample cases only. The weight used in that plot is the original XS sample weight. The lower left-hand corner presents a similar graph comparing weighted and unweighted AGI distributions for all 1983 SCF cases reporting 1982 AGI of less than $100K. The weighted distribution for this subplot is based on the composite weight. The right-hand side of Figure 3 provides a similar comparison of distributions for sample cases reporting 1982 AGI above $100K. The subplot in the upper right focuses only on HY sample cases; the weighted distribution is estimated using the original IRS HY sample weight. The companion subplot in the lower right includes both HY and XS sample cases with $100K+ AGI to compare the composite weighted distribution to the unweighted AGI distribution for these sample cases.

The impact of weighting on the distribution of AGI in the less than $100K range is almost unnoticeable at the level of resolution provided by the graphs. For the AGI range above $100K, the effect of the weighting on the cumulative distribution function is significant. Since the IRS HY sample is known to be disproportionately allocated to strata of tax filers with higher incomes, the HY sample weight should increase with decreasing income. Relative to the unweighted distribution of AGI, it is then intuitive that the weighted sample distribution of AGI will be shifted toward the graph's X-origin of $100,000—the amount of the shift being a function of the degree to which reported AGI and the HY weights (or assigned composite weight values) are negatively correlated. The subplots on the right-hand side of Figure 3 clearly illustrate the expected distributional shift that occurs when weights are applied to the AGI measure from the survey.

The original composite weight variable for the 1983 SCF was developed in accordance with the known design-based selection and response properties of the data set. However, because of the complexity and uncertainty introduced by the special sampling of high income taxpayers, a number of assumptions and approximations were required to construct a composite weight variable for joint analysis of the XS and HY sample data sets. Clearly, the focus of real concern over the weighted analysis of the 1983 SCF data must be on the upper income tail of the sample distribution. Above the arbitrary $100K AGI cutoff point, the income distribution (i.e., income

FIGURE 3.--CUMULATIVE SAMPLE DISTRIBUTIONS OF AGI



estimates) is more sensitive to the choice of a weighting adjustment.

V.B. Improvements to the Original 1983 SCF Composite Weight

During the time period following the computation of the original composite weight for joint analysis of the 1983 SCF XS and HY sample data, the IRS has provided additional detail related to the selection of the HY sample. Specifically, the IRS has provided SOI frame counts of the number of tax filers in each of nine HY-sample strata. For each HY sample stratum, population counts were further disaggregated according to the self-representing/nonself-representing status of the primary stage sampling unit (PSU) of the taxpayer. While these additional data contribute little toward a better understanding of the stratification itself or the "boundary" of the HY sample frame, they do provide valuable information on the total size of the HY sample eligible population.

In revising the original composite weight, the first step was to introduce two corrections to the weight components:

1) Weight values for HY sample cases were scaled so that the sum of weights in the self-representing and nonself-representing divisions of each of the nine HY sample strata matched the stratum control totals provided by the IRS.
2) Weight values for XS sample cases received an additional nonresponse correction to compensate for the post-survey deletion of n=159 XS sample cases which were judged to be too incomplete or unreliable to be used in analysis.

Following these corrections a preliminary version of the revised composite weight was computed using an algorithm similar to that outlined in the discussion of the original composite weight. Next, each sample case was assigned to a cell of the four by two matrix representing a cross-classification of Census Region and AGI range (<100K, 100K+). Iterative fitting or "raking ratio estimation" was then applied to align the marginal sums of cells' aggregate weights to 1 July 1983 household totals for the four Census regions and to specified control totals for households by AGI bracket. For the latter, the 1983 count of households in the $100K+ range was set equal to 706,000 households—the 1980 IRS population count for tax filers eligible for the HY sample.

V.C. Comparing Income and Wealth Estimation Properties of the Original and Revised Composite Weight Factors

In analysis, the objective in using the original composite weight or the revised version of that weight is to produce estimates with a minimum of mean square error (MSE=Variance + Bias^2). Since both composite weights include substantial adjustments for nonresponse and approximate (as opposed to exact) post-stratification corrections for noncoverage or sampling departures from population distribution controls, it is unlikely that either will produce income and wealth estimates that are completely free of bias. Intuitively, the added HY sample post-stratification controls favor the revised sample weight. Unfortunately the true residual bias associated with the use of these weights is impossible to measure. Consequently, it is difficult to say which weight will yield estimates with lower mean square error.

Although the bias component of mean square error cannot be reliably measured, the sampling variation of estimates computed using the two weights can be compared. A direct comparison of estimates also is useful for determining the sensitivity of sample statistics to the two composite weighting alternatives. Table 4 provides a comparison of estimates of mean AGI and their standard errors computed using 1) the original and 2) the revised composite weight variables.

In the lower AGI ranges, the choice of a composite weight appears to have little effect on either the value of the estimated mean or its standard error. This is to be expected since the major adjustments in the revised composite weight operate on cases in the higher income HY sample group. Interestingly, standard errors of estimates computed using the revised composite weight tend to be slightly higher than those of estimates derived using the original composite weight. In Section VI, we will show that the observed increase in standard errors can be linked to increased "weighting effects" of the revised composite weight variable.

Table 4.--Comparison of Mean AGI Estimates and Their Standard Errors Using 1) the Original and 2) the Revised Composite Weighting Factors

AGI Subclass Range	With Original Composite Weight		With Revised Composite Weight	
	Estimated Mean	Standard Error	Estimated Mean	Standard Error
Total	$27,660	$729	$25,030	$775
<$25K	$12,424	$195	$12,430	$195
<$50K	$19,428	$416	$19,402	$418
<$100K	$23,063	$542	$23,041	$546
<$200K	$24,563	$572	$24,567	$579
<$500K	$26,469	$644	$24,427	$680
>$50K	$101,285	$2,815	$85,651	$3,490
>$100K	$241,137	$8,547	$267,131	$16,258
>$200K	$397,569	$14,887	$401,769	$28,149
$50-99K	$64,991	$808	$65,096	$838
$100-199K	$128,315	$2,291	$137,608	$1,478
$200-499K	$294,398	$5,039	$292,030	$5,972

VI. SAMPLING ERRORS OF 1983 SCF INCOME ESTIMATES

Sampling errors of estimates based on data collected under the complex sample design of the 1983 SCF survey are influenced by: 1) the population variance of the income characteristic(s) on which the estimate is based; 2) the effectiveness of sample stratification; 3) the degree of clustering of sample elements; and 4) effects of non-optimal weighting of the sample observations.

Tables 5 and 6 present two sets of sampling error results for 1983 SCF estimates of mean household AGI. Results of sampling error computations are described for the total population and subclasses defined by selected cumulative ranges, closed interval ranges and open ended classes of respondents' 1982 AGI. Table 5 estimates are computed using the original SCF composite weight. Table 6 estimates are based on the revised composite weight..

Columns one and two of each table identify the income subclass for which the mean is being estimated. The third column of each table provides the sample size base for the 1983 SCF estimate. The fourth through sixth columns provide the estimated mean value, its standard error estimate, and the corresponding coefficient of variation for the estimate, $cv(\bar{y})=se(\bar{y})/\bar{y}$. The column labeled DEFT contains estimates of the square root of the "sample design effect" for the estimated mean AGI statistic. The sample design effect measures the precision of the complex sample design relative to that obtained from a simple random sample (SRS) of equivalent size.

DEFT reflects the combined effect of sample design stratification, clustering and weighting on the standard error of estimated means. For example, a value of DEFT=1.10

Table 5

Standard Errors of 1983 SCF Estimates:
Mean Value of Household Income
(Original 1983 SCF Composite Weight)

Total Population	Income Range	n	Estimated Mean	Standard Error	CV	DEFT	L*
		4103	$27,660	$729	.026	.963	4.224
Cumulative Range	<5K	358	$3,306	$81	.025	.847	1.022
	<10K	916	$5,773	$116	.020	1.388	1.007
	<25K	2283	$12,424	$195	.016	1.434	1.001
	<50K	3328	$19,428	$416	.021	1.965	1.001
	<100K	3632	$23,063	$542	.024	1.873	1.034
	<200K	3814	$24,563	$572	.023	1.634	1.431
	<500K	4004	$26,469	$644	.024	1.281	2.221
	<1M	4066	$27,151	$713	.026	1.185	2.917
Selected Closed Intervals	5-7.5K	299	$6,109	$54	.009	1.234	1.002
	7.5-10K	259	$8,712	$50	.006	1.183	1.005
	10-15K	522	$12,426	$69	.006	1.091	1.000
	15-20K	461	$17,265	$61	.004	0.890	1.008
	20-25K	384	$22,138	$71	.003	0.972	1.004
	25-30K	322	$27,068	$82	.003	1.006	1.006
	30-40K	465	$34,296	$129	.004	0.953	1.001
	40-50K	258	$44,082	$186	.004	1.025	1.012
	50-99K	304	$64,991	$808	.012	1.059	1.144
	100-199K	182	$128,315	$2,291	.018	1.133	1.040
	200-499K	190	$294,398	$5,039	.017	0.932	1.101
Open Intervals	>50K	759	$101,285	$2,815	.026	0.568	3.433
	>100K	458	$241,137	$8,547	.035	0.783	2.151
	>200K	283	$397,569	$14,887	.037	0.822	1.851
	>500K	96	**	**	**	**	**

* Loss factor indicating increase in standard error due to weighting.
** Estimates of variance for this open-ended category proved unstable.

Table 6

Standard Errors of 1983 SCF Estimates:
Mean Value of Household Income
(Revised 1983 SCF Composite Weight)

Total Population	Income Range	n	Estimated Mean	Standard Error	CV	DEFT	L*
		4103	$25,030	$775	.031	1.368	5.630
Cumulative Range	<5K	358	$3,305	$80	.024	0.838	1.027
	<10K	916	$5,775	$114	.020	1.363	1.008
	<25K	2283	$12,430	$194	.016	1.429	1.001
	<50K	3328	$19,402	$418	.022	1.981	1.002
	<100K	3632	$23,041	$546	.024	1.887	1.033
	<200K	3814	$24,567	$579	.024	1.870	1.622
	500K	4004	$24,427	$680	.028	1.738	2.873
	<1M	4066	$24,773	$745	.030	1.638	3.882
Selected Closed Intervals	5-7.5K	299	$6,110	$55	.009	1.246	1.000
	7.5-10K	259	$8,710	$50	.006	1.187	1.000
	10-15K	522	$12,426	$69	.006	1.089	1.002
	15-20K	461	$17,262	$61	.004	0.887	1.000
	20-25K	384	$22,136	$71	.003	0.979	1.004
	25-30K	322	$27,070	$81	.003	0.998	1.006
	30-40K	465	$34,290	$129	.004	0.954	1.000
	40-50K	258	$44,074	$185	.004	1.018	1.013
	50-99K	304	$65,096	$838	.013	1.090	1.071
	100-199K	182	$137,608	$1,478	.011	.755	1.018
	200-499K	190	$292,030	$5,972	.020	1.085	1.084
Open Intervals	>50K	759	$85,561	$3,490	.041	0.951	4.423
	>100K	458	$267,131	$16,258	.061	1.337	1.894
	>200K	283	$401,769	$28,149	.070	1.457	1.730
	>500K	96	$895,677	$55,321	.062	1.038	1.294

*Loss factor indicating increase in standard error due to weighting.

implies that the design stratification, clustering and weighting combine to produce a 10% increase in standard error relative to the standard error of the mean expected from an SRS sample of equal size. Effective stratification and sample allocation to strata will operate to reduce the value of DEFT. Clustered sampling and the associated intraclass correlation among cluster elements produce increases in the variance of sample elements with a corresponding increase in the DEFT. Random or otherwise non-optimal weighting of sample cases also leads to higher values of DEFT.

Examining Table 5 and 6, DEFT values for estimated means of AGI are highly sensitive to the income range for which the mean value is being computed. For closed interval ranges of the AGI variables, DEFT values tend to be slightly greater than 1.0 — very modest clustering and weighting effects are present. For interval ranges bounded from above, DEFT values are larger than those for closed interval estimates.

The inflation of variances caused by the need to use weights to compute the 1983 SCF sample estimates is confounded with stratification and clustering influences in the sample design effect, however, through an alternative computation the weighting effect can be separately estimated. The final column of Tables 5 and 6 provides estimates of the precision loss factor, L, — where L-1 represents the proportion by which weighting increases the standard errors of the estimated mean values relative to an unweighted sample of similar design and sample size.

VII. SUMMARY

The 1983 Survey of Consumer Finances provides an excellent illustration of many of the difficulties and problems associated with survey-based research on income and wealth characteristics of the general population. Through its dual frame sample design, the 1983 SCF has moved in the direction of addressing the issue of optimal allocation of sample observations—that is, placing more than a proportionate share of observations in the upper income strata of the distribution where income and wealth characteristics exhibit the highest levels of variation.

While implementation of the dual frame design provides an avenue for addressing theoretical concerns over sample allocation, many practical problems remain. Access to the list frame used in the 1983 SCF had to be tightly controlled by the IRS. As a consequence, little is known about the true nature of the stratified sampling design. The absence of complete documentation for the HY sample complicates the dual frame estimation procedures. Even the development of a general multi-purpose composite weight factor is hampered by the uncertainty over strata definitions and stratum specific sampling rates. Differences in the reference periods of the survey (1982 tax year) and the SOI frame used for the HY sample (1980 tax year) also complicate weighting and estimation procedures.

The taxpayer informed consent requirement of the HY sample selection process resulted in very low response rates among high income sample individuals. Data which would provide some evidence on the nature of nonresponse bias in the HY sample are also not yet available, although the IRS is currently conducting its own study of nonresponse bias in the HY sample data set.

With so many uncertainties and problems, why bother with the added complication of the dual frame design? Possibly, the simplest answer is that we have no other choice if we want to continue to study income and wealth characteristics of the general population—particularly if there is interest in households in the upper income ranges. (Standard area probability sample designs such as those used in the CPS and SIPP should suffice for the study of lower and middle income groups.) Despite the many assumptions and approximations which have gone into preparing the survey data for analysis, 1983 SCF provides an extremely rich source of data for the analysis of relationships among income, assets and total wealth (net worth).

Presently, SRC is conducting additional research into weighting alternatives for the 1983 SCF. The outcome of this work is important not only to analysts of the 1983 data but also for researchers who will be working with data from the 1986 re-interview of 1983 SCF respondents.

For the future, the 1983 SCF has provided both methodological experience and an information base which allows researchers to refine the statistical and methodological features of the dual frame approach to income surveys. Beyond the immediate concern over the difficult weighting and estimation issues that have been discussed at some length in this paper, questions related to the handling of outlier values, adjustments for unit nonresponse, and imputation of item missing data also require further research and methodological development.

FOOTNOTES

[1]The adjective "design-based" is used to imply methods of estimation and inference which draw on the sampling distribution properties of the survey estimate under the given probability sample design.

[2]SMSA=Standard Metropolitan Statistical Area.

[3]There is also strong reaction from researchers to the prospect of throwing away data for some analyses.

[4]The use of the common term "pooled analysis" is avoided here since it carries a quite different connotation in experimental statistics.

[5]PSU=Primary Sample Unit.

REFERENCES

Cochran, W. G. (1963). Sampling Techniques, Second Edition. John Wiley & Sons. New York, NY.

Kish, L. (1965). Survey Sampling. John Wiley & Sons. New York, NY.

Dalenius, T. (1957). Sampling in Sweden. Almqvist & Wiksell. Stockholm, Sweden.

Hartley, H.O. (1962). "Multiple Frame Surveys," Proceedings of the Social Statistics Section, American Statistical Association, pp. 203-206.