The Good Shepherd:
Sample Design and Control for Wealth Measurement in the Survey of Consumer Finances

Arthur B. Kennickell
Senior Economist and Project Director Survey of Consumer Finances
Board of Governors of the Federal Reserve System
Mail Stop 153
Washington, DC 20551 USA
Email: Arthur.Kennickell@frb.gov

January 2005

Presented at the January 2005 Luxembourg Wealth Study Conference
Perugia, Italy

 The author wishes to thank the many talented colleagues with whom he has developed his understanding of technical aspects of measurement, particularly Martin Frankel, Steven Heeringa, Roderick Little, Donald Rubin, Fritz Scheuren and Louise Woodburn.

I. Introduction

This paper focuses on the sample design issues related to the collection of wealth data in the U.S. Survey of Consumer Finances (SCF). Sample surveys are complex "measurement engines" that may be viewed as falling into three important parts: selection and pursuit of participants, collection of data, and statistical processing of the resulting information. The sample design provides the most fundamental measurable statistical basis for it all. It is obvious that a good design should provide the most efficient and unbiased representation of the population relevant for the measurement task at hand that is feasible with the available resources. At the same time, however, a design that is well integrated with the objectives of a survey may also have implications for the management of data collection and post-survey processing. Indeed, a survey that fails to capture at least some such benefits must either be inefficient or highly constrained.

Sampling is a structured means of selecting units from a *population of interest* in order to represent that population in terms of a set of *characteristics of interest*. The act of sampling requires a statistical *design* appropriate to the entire measurement task. To be maximally useful, the design must reflect formal mathematical constraints, encompass as well as possible behavioral and logistical factors implicit in pursuing and obtaining data from the units selected, and provide a framework for post-survey adjustments. In setting a design for the collection of wealth data, four factors are particularly important.

First, because the distribution of wealth is skewed, relatively large shares of total wealth are held by relatively small parts of the population. Using data from a purely random selection of units, for example, would at best yield a statistically very inefficient estimate of the distribution of wealth. Second, a variety of factors may influence the willingness of the selected units to cooperate: a sense of limited time, a strong sense of privacy, suspicion of the data collector, social alienation, etc. If units vary in their cooperativeness and that variation is also correlated with wealth, as experience suggests, then measurement bias will result. Third, samples are most often implemented by interviewers, who face a set of behavioral incentives. Typically an overwhelming consideration in retaining and rewarding interviewers is their production of completed interviews. Unless there is a means of controlling the implementation

of a sample, rational interviewers who faced such a system would be more likely to exert most effort on cases most likely to be completed, thus exacerbating any biases resulting from decisions made by the sample members. Finally, depending on the structure of the sample, there may be factors correlated with aspects of wealth that can be observed for a given design, either based on register data of some sort or on observations made on all sample elements during the time when the sample is implemented. Where a sample can be chosen in a way that aligns with such information, that information may be used to guide post-survey adjustments to compensate for nonresponse and possibly to reduce sampling error. Each of these points will be developed further in this paper.

The next section of the paper provides a brief and informal overview of the a few essential technical issues in sample design. The subsequent section describes the sampling approach used in the SCF. Section IV addresses problems of sample control during the field period and Section V reviews possibilities for post-survey adjustments to increase efficiency or reduce bias. The final section of the paper summarizes the most important points and makes suggestions for progress in this area.

II. An Overview of Sampling

As an ideal, analysts would like to have survey data that form a microcosm of the population relevant for their research that could be used to address every question with perfect accuracy. However, it is only in censuses where participation is complete that there is even such a possibility. Censuses are most often prohibitively expensive in several dimensions, and the feasible amount of detail that can be collected is usually small. This painful reality long ago drove researchers to consider selection of samples. After much creative groping toward a scientific approach, the seminal paper of Neyman [1934] appeared and established the foundation of modern probability sampling.

From Neyman's paper, sampling evolved into one of the most elegant areas of mathematical statistics and a key tool in scientific data collection. Although there is a vast sampling literature, there are two scholarly works that may be taken to span the essential technical material: Kish [1965] and Särndal, Swensson and Wretman [1992]. In both, sample-

based estimates are treated as realizations of random variables arising from some process in the relevant population. In the purest version of the former work, random sampling without explicit model assumptions is shown to guarantee for a broad class of estimators (such as the mean) that if estimates are made repeatedly from many identically structured samples, the average of all estimates converges to the true value. In addition, this approach gives a way of characterizing the probability conditional on the sampling process that the true value might actually be within an interval some distance from the estimate given by a particular sample.

But in a given sample, we usually have no way of knowing whether that sample is, in fact, one yielding an estimate far from the true value or not. Knowing that the probability of selecting such a sample is very small–and that, consequently, the probability of the estimate being distant from the true value is small–is cold comfort. Thus, many of the early extensions focused on techniques that might reduce the inherent variability of estimates both within a given survey and in repeated measurements over time. Those extensions incorporated some assumptions about population structures, but still framed within the context of replication. Some of the later work began to make assumptions that rely more heavily on explicit models governing characteristics of the relevant population. Economists, a generally model-dependent group, would often feel at home with such assumptions. The relationship between the replication framework and the model-based framework remains controversial in statistics. In practice, however, samplers of different persuasions often make comparable decisions, even though their underlying motivations may differ.

Two technical points are worth presenting in a small amount of detail here as motivation for what follows: the distribution of estimates under simple random sampling and the effect of a particular method of imposing additional structure on a sample. Much of the sampling literature deals with the effects of different types of sampling on estimation of the mean of the distribution of a given variable *y*. Under simple random sampling with sample size *n* from a population of size *N*, the estimate of the mean is given by:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The standard error of the estimate of the mean is given by:

$$\left(1-\frac{n}{N}\right)^{\frac{1}{2}}\frac{\sigma_y}{\sqrt{n}}$$

where σ_y is the standard deviation of the distribution of *y* in the full population. In the usual case where *n* is small relative to *N*, the first term is negligible. There are two important points to note. First, the variability of the estimate depends on the variability of the variable. Second, the decrease in the variability of the estimate is concave in the size of the sample. That is, there are decreasing returns from increasing sample size. Because costs of increased sample size are usually closer to linear, there will be an optimal point where the value of a decrease in variability exceeds the marginal cost of the reduction. Although these formulas apply only to estimates of means, similar intuition applies generally to a broad class of other estimators.

It follows from the formula for the standard error of the mean that populations with widely varying values of a variable of interest will have a relatively broad distribution of sampling error. A special subclass of variables that exhibit high variability is the case where there is a relatively small group that possesses a quality not generally possessed by other groups. As a simple example, consider a population where 99 percent of units have a characteristic *y*=1, the remaining 1 percent have a value of *y*=1,000 and the estimate of interest is the mean of *y*. The "rare" quality here is having a large value of *y*. If one repeatedly applied simple random sampling and estimated the mean of *y*, a distribution of means like the simulated one given by the solid line in figure 1 would result.[1] The true value of the mean is 10.99, which is both the mean and the median of the empirical distribution shown. But the values in many not-too-unlikely samples differ substantially. As shown in figure 2, well over 20 percent of possible samples have errors of over 10 percent. This example captures in spirit the pure sampling problem of measuring wealth when the distribution is highly skewed, as it is in most countries.

Where there are groups in the population that either posses a rare trait of interest or that exhibit relatively high variability of the variable of interest, there may be gains from sampling disproportionately larger fraction of observations from those groups. Even where variability is

[1]The distribution is simulated using 10,000 samples of 10,000 observations each.

not as extreme as in the example, there may be returns from enforcing some key distributions in the population in the selection of the sample. Such sampling is called "stratified." In the example, if one could identify the rare group precisely as a sample stratum, the estimated distribution would always produce the mean exactly. More realistic are instances where groups can only be identified approximately, but the intuition is similar. There are formulas that may be applied to derive an optimal size for sample strata in many cases. As noted later in this paper, stratification may also be helpful in post-survey adjustments in reducing nonresponse bias.

A particularly important application of stratification in household surveys is in the creation of geographically clustered samples. In surveys that require in-person visits by interviewers, there are often great cost savings by sampling in such a way as to minimize the distance between the locations of households, at least within clusters. A key sample of this type is the multi-stage area-probability design. In the most common such designs, selection operates through several steps to yield a sample where every unit ultimately selected has an equal probability of selection. Large geographic units are classified into strata using an array of characteristics, and areas are selected with a probability proportionate to the size of the population of the groups. Within each of the selected groups, sub-areas are stratified and as at the first stage, areas are selected with probability proportionate to population. This second step may be repeated several times to reach a stage where relatively small "neighborhoods" have been selected. Within those neighborhoods, there is typically no stratification; every household is listed and a random selection is made. Area-probability designs tend to be very robust for measurement of qualities that are spread broadly throughout the population.

The concepts of randomization and variability deserve a final emphasis in this section. The function of randomization, at whatever level it may be applied, is to reduce the risk that "judgment" or other systematic factors in the selection of observations might yield false conclusions. At the same time, it provides a mathematical apparatus to characterize the latent variability of outcomes. To speak of sample-based estimates without reference to a measure of associated variability is to discard a very large part of the scientific apparatus of sampling, and in some cases to be misleading. A corollary of this argument in surveys of wealth is that having an estimate of total wealth close to independent aggregates is not particularly meaningful unless the variability of the survey estimate is relatively small.

Figure 1: Distribution of Mean(y) under simple random sampling.

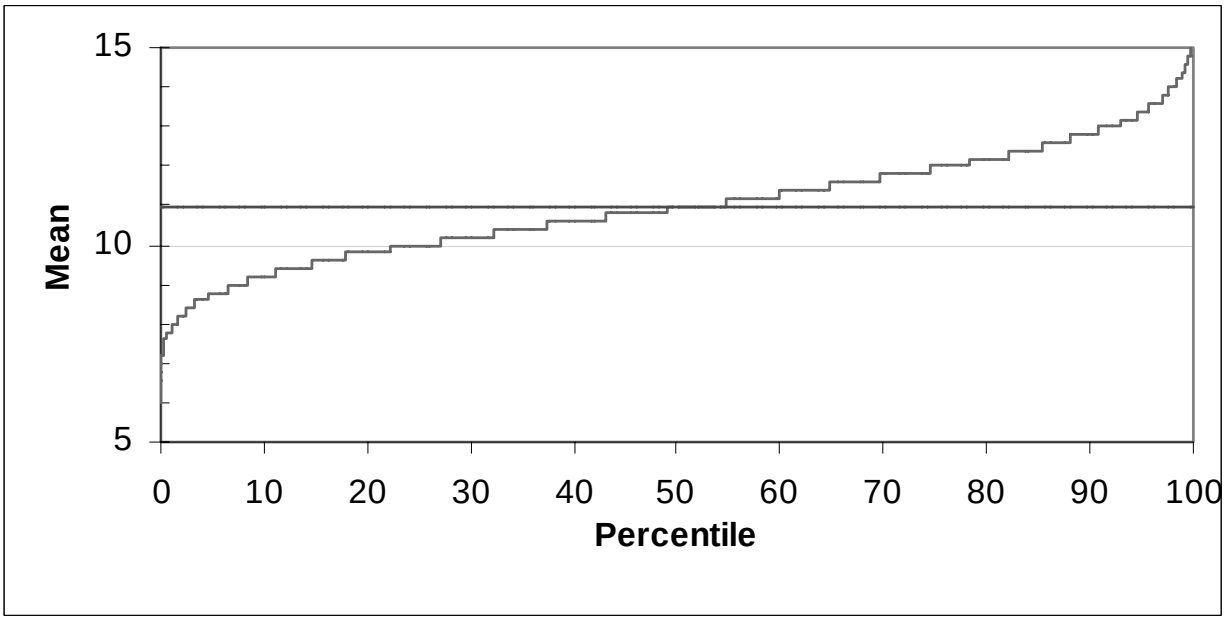
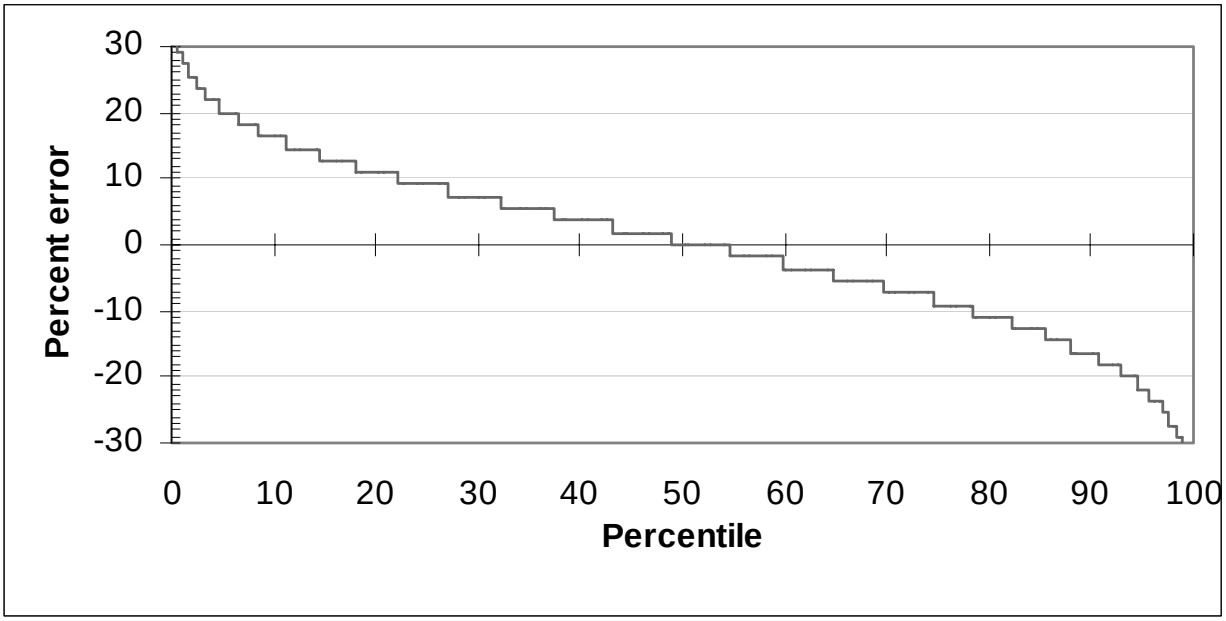


Figure 2: Distribution of percent error in Mean(y) under simple random sampling.



III. Sampling in the SCF

The SCF is required to provide estimates of characteristics of relatively rare variables–such as direct holdings of corporate bonds–and other more common variables–such as holdings of owned principal residences.[2] A standard area-probability sample of the sort sketched earlier is used to give adequate representation of the common variables and to provide a basis of reference for the more narrowly held variables. A special supplemental sample is stratified to give a more efficient representation of the narrowly held variables; most of the remaining discussion in this section focuses on this sample.

In a situation where only the estimation of wealth, not components of wealth, was the overwhelming goal, the ideal supplemental sample would be so tailored that units would be selected from different wealth groups to minimize the variability of the key statistics needed. Although the demands placed on the SCF are far broader, this narrow view serves as an adequate approximation for expository purposes.

The supplemental sample for the SCF is a type of "list" sample–that is, a sample selected from a list of individual units, rather than one "discovered" through a mechanism as in area-probability sampling. The list that serves as a basis of the sample is a file of statistical records derived from individual tax returns by staff at the Statistics of Income Division (SOI) of the U.S. Internal Revenue Service.[3] Under a legal contract, part of this sensitive information is shared with the project staff of the SCF at the Federal Reserve Board exclusively for purposes of design, execution and processing of the survey. The file contains a subset of the monetary items that appear on an individual return supplemented by filing status, birth date of the filers, and other such variables.

---

[2]See Aizcorbe, Kennickell and Moore [2003] for a discussion of the SCF data and Kennickell [2000] for an overview of the technical background to the survey.

[3]The SOI file is based on a sample selected from all federal individual income tax returns filed within a given year. Because the file samples taxpayers with high incomes or unusual characteristics at a very high rate, it is a sufficient basis for the SCF sample. The SOI data are specially edited to resolve irregularities. See Internal Revenue Service [2001] for a description of the SOI data and the selection process for that sample.

There are four central technical problems with using such information for SCF sampling.[4] First, the information included in the SOI file at best addresses flows of income from assets, not the assets themselves. As noted in more detail below, this restriction leads in the SCF to the use of models to connect income with wealth.

Second, the connection between the observed income flows and the underlying assets may reflect rates of return that vary widely across individuals–depending on skill, opportunity, risk preferences and luck. Where sufficient information is available, it may be possible to use models to account for these differences.

Third, not all the income flows that are relevant in estimating wealth necessarily appear directly on a tax return. Many lawyers and accountants earn a living finding legal means of minimizing the amount of income realized in a taxable form. Such problems may be especially severe in the case of personal businesses where the receipt of income may legally and more easily be deferred or manipulated to minimize tax obligations. Some assets are structured to yield very little current income, but to cumulate value in terms of capital gains that may be realized in some period. In some cases legal structures, such as trusts, may be used to hold assets. Such arrangements may or may not generate reportable personal income; the wealth contained may or may not be recoverable by a person who is a beneficiary under the arrangement; in some cases, such as charitable trusts, the person who established the trust may retain full control of the assets for a restricted set of purposes; many other such complications are possible, including less formal family arrangements. If wealth is defined as net worth, rather than gross assets, there may be problems because many categories of liabilities may be missing from a return altogether. These are serious problems for straightforward translation of income into wealth, but again, modeling offers some hope of mitigating them.

Fourth, the unit of observation in the SOI data is a tax-filing unit, which may be a single individual, a married individual (or one partnered without benefit of marriage) filing a separate return, or a a couple filing a joint return. Households, the unit of interest in the SCF, may be much more complicated. Among the simpler situations, a given household may contain multiple

---

[4]See Kennickell and McManus [1993] for a more detailed review of technical problems in sampling from SOI data.

people who file tax returns, but the available data make it impossible to account fully for such linkages. In some households, no one may file a tax return, because an income level below the filing threshold or other considerations obviate having to do so, because there is illegal income not reported to the tax authority, or because of error or wilful defiance of the law. In the SCF, the primary need for the SOI data is in targeting people who are disproportionately likely to be wealthy. In practice, most problems of the fourth type have their greatest effect at the opposite end of the wealth spectrum. Of issues related to the definition of the household unit, the only one that is particularly awkward is changes in marital status since the year the tax return that anchors the sample was filed. When there is a separation or divorced of a couple by the time a member of the original unit is approached by an interviewer, an attempt is made to interview both people separately. New marriages raise similar problems. The frequency of such unit changes in the higher strata of the list sample is relatively low, and by now the cumulative SCF experience is that the statistical adjustments made to accommodate the marital status changes are not obviously distorting.

Clearly modeling is key in translating the SOI income data into systematic information more closely tied to wealth for sampling.[5] For the SCF, two classes of models are used to predict a "wealth index," which is then taken as an approximate instrument for ranking taxpayers in terms of their wealth for stratified sampling. In the simplest such model, at time t for case i, every asset A_{ijt} has a rate of return r_{jt} yielding income y_{ijt}, so that total wealth is given by

$$WINDEX0_{it} = \sum_j \frac{y_{ijt}}{r_{jt}}.$$

A model of this type was used in the design of the SCF list sample in 1989, and it has been an element of the design since then. The income figures used are the capital income amounts that appear on an individual income tax return: taxable and nontaxable interest, dividends, business income, and capital gains. The rates of return are average rates applying during the time the income was generated. Although it is unrealistic to expect that rates of return are constant both

[5]See Kennickell [1999a and 2001] for more detailed discussion of modeling wealth in terms of income in the SCF sample design.

across individuals and across the varieties of assets potentially underlying each income type, the approach has the advantage that rates of return enter the model transparently. Furthermore, it seems not unreasonable to think that in a long-run average sense, the structure should be adequate. For individual cases and classes of cases in the short term, however, the result may be quite different.

An alternative approach is to define wealth as a more complex function of income and other characteristics affecting differential rates of return, propensities to have assets with nontaxable returns, etc. as follows:

$$WINDEX1_{it} = \Gamma\left(\rho\left(X_{it}, Y_{it}\right)' Y_{it}, X_{it}, Y_{it}\right),$$

Where $\rho(X_{it}, Y_{it})$ is a vector of the inverses of the rates of return for Y_{it} and where X_{it} and Y_{it} are included separately to model unreported income, "missing" household members, and other conceptual and practical imperfections of the sort discussed above that are not accounted for in the more straightforward WINDEX0 model. In practice, many of the components of X_{it} that are potentially important in this model are not available, so the variables that are available must serve as proxies. Unlike the straightforward model, this one must be estimated; that is, survey data and tax-based data must be combined directly.

Following a very long negotiation, permission was gained during the preparation for the 1995 SCF to perform a match of the 1992 survey wealth data with the single year of SOI data that had been used in selecting the 1992 sample.[6] A variety of specifications were tried to find the best fit of the wealth data, consistent with the inclusion of a core set of variables and a few general theoretical principles. The data strongly rejected the WINDEX0 income-return model as the optimal specification, and coefficients estimated for the narrow set of variables in that model implied unbelievable rates of return.

One serious caution about the estimated model is that the parameters are implicitly a function of the rates of return in 1992 and potentially other period-specific factors. Thus, using

[6]As usual in SCF work with such sensitive information, access to this linked information was severely limited and the file used for the analysis was stripped of case identification numbers.

those coefficients to predict wealth for the 1995 using income data from three years later risks misclassification simply as a result of changes in the embedded rates of return and other period effects. Unlike the case in the WINDEX0 model, there is no obvious way to intervene directly to adjust to estimated coefficients to offset changes in returns.

The two modeling approaches present different risks of classification errors. Factors favorable to the WINDEX0 model are its transparency and its use of period-specific rates of return, but that model ignores much other complexity, particularly idiosyncratic variation in rates of return. The WINDEX1 model offers the possibility of accounting for varying rates of return and greater complexity in the relationship between wealth and returns, but it risks error by its inability to account for changes in the structure of returns and other period effects. As in other situations where there are competing imperfectly correlated measures with different risk properties, pooling is an attractive option.

For the 1995 survey, each of the two indices was computed for all cases in the SOI file of data available for 1993.[7] Each series was then standardized by subtracting its median and dividing by its interquartile range, these adjustment factors being more robustly estimated than the mean and standard deviation of the predicted distributions. The pooled estimate, WINDEXM, was the simple average (in the absence of strong information to derive optimal pooling weights) of the standardized values of WINDEX0 and WINDEX1 for each observation. Strata were defined in terms of the percentiles of the distribution of WINDEXM, and the break points of the strata were chosen to maintain the approximate fraction of the full population of taxpayers in each group as under the 1992 definitions based on the WINDEX0 model.

Since the 1995 survey, there have been two main lines of development in the list sample design. First, the WINDEX1 model has been reestimated with each succeeding wave of data and progress has been made in terms of its predictive ability. Informal evidence accumulated suggests that the improvement has been substantial; formal evidence will be available in a later paper.

---

[7]These SOI data very largely consist of data based on tax returns filed in 1994 for 1993 income.

Second, as a help in reducing the classification error in the models due to temporary deviations from a more permanent income, additional years of tax-based data have been made available for the estimation of the WINDEX1 model and the simulation of WINDEXM on multiple years of later data for sampling. The income data may be variable for many reasons. For some people, changes in employment status are the most important driver of income changes, for others changes in family structure–birth, death, divorce, etc.–are important, for some volatility of the financial markets or particular business markets is key, while for others tax considerations are more important in determining realized income. Because income is typically more volatile than wealth, sampling on the basis of one year of income data would tend to yield a noisier estimate of wealth than if a measure more like permanent income could be used. Figure 3 shows selected percentiles of the distribution of the coefficient of variation (ratio of the observation-specific standard error to the observation-specific mean) of total taxable income for tax years 2000 through 2002, where incomes for each taxpayer for 2001 and 2002 have been standardized to 2000 using the aggregate rate of growth for each year.[8] The horizontal axis organizes the data by the stratum of the list sample in which each observation was contained. As is clear from the figure, there is substantial variability even at the median, where the coefficient of variation is roughly 8 percent; at the 90th percentile it exceeds 40 percent overall. Moreover, variability tends to be greater among taxpayers who would be in the higher strata of the SCF list sample. Thus, one would expect a large return from incorporating multiple years of income data into the sampling model.

[8]The estimates shown in the figure income only observations that were included in the data file for all three years. Some taxpayers were excluded because of a meaningful change in their filing status (e.g., from married filing jointly to married filing separately).

Figure 3: Distribution of the coefficient of variation of standardized total taxable income 2000-2002; 25th, 50th, 75th and 90th percentiles of the distribution; by list sample stratum.

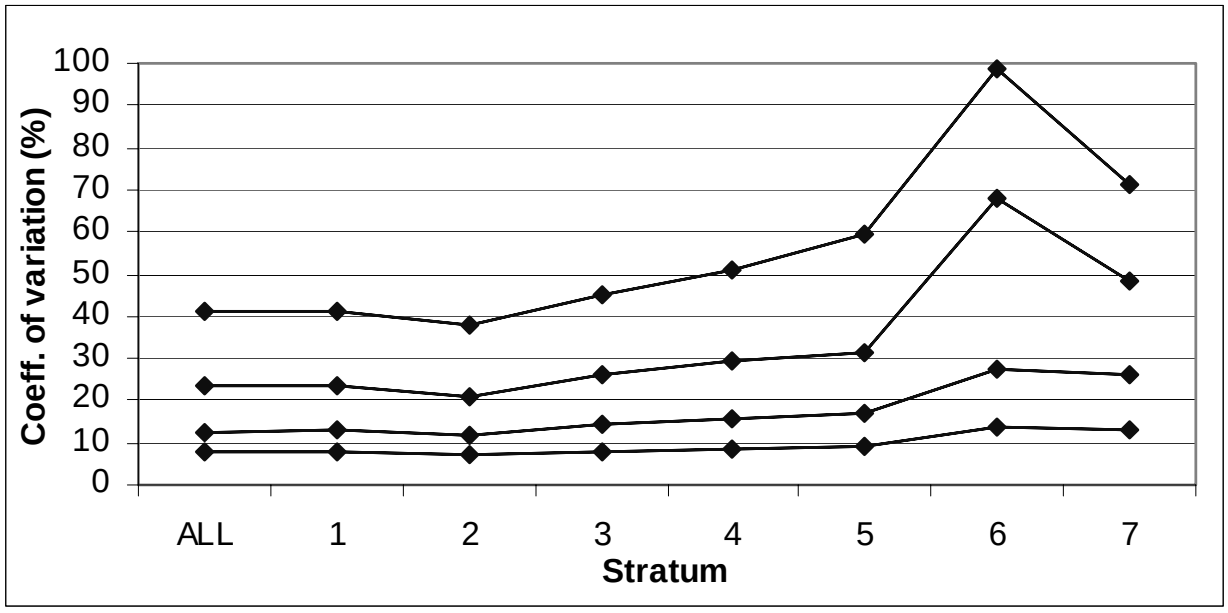
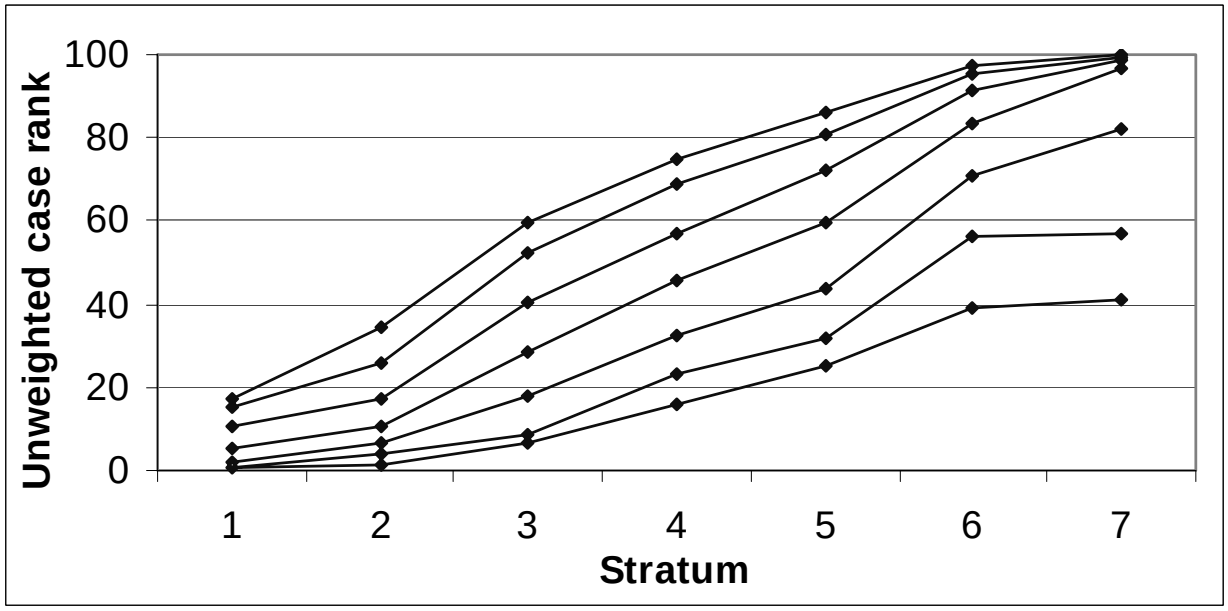


Figure 4: Distribution in the rank of net worth in the 2001 SCF list sample; 10th, 25th, 50th, 75th and 90th percentiles of the distribution; by list sample stratum.



In the 2001 SCF, two years of tax-based data were used to estimate the WINDEX1 model and to predict both WINDEX0 and WINDEX1.[9] For the 2004 SCF, three years of income were available. A later paper will investigate the importance of the gains from the use of the additional information. However, it is possible to characterize generally the extent that the list sample design achieves its primary purpose of classifying taxpayers by their wealth for sampling. Figure 4 shows the distribution of the rank of 2001 SCF list sample cases interviewed (the most recent data available for this analysis) in terms of net worth across the sample strata. Although there is a non-negligible amount of "misclassification," the list sample design overall provides good separation of taxpayers by wealth.

IV. Sample management

Great care is usually devoted to ensuring that a sample design is as close to unbiased as possible and that its efficiency properties are optimized for the measurement task. But the sample specified is almost always larger than the part of the sample for which data are ultimately obtained. Typically, complex factors key in the implementation a design lead to a type of implicit subsampling that is not directly under the control of the sampler. As much of these factors as possible should, in theory, be taken into account in the *total sampling process*.

One set of problems concerns the validity of units selected into the sample. Often, some respondents cannot be identified or located. Even in geographically-based samples, some units cannot be located. Moreover, sometimes pre-existing units have been demolished, converted to other uses or replaced with new construction, possibly of an entirely different scale or form; in name-based samples the person may be deceased, or where couples are the named sample element, there may have been a divorce. When respondents cannot be located, it is uncertain

[9]Ideally, one would link successive years of SOI data, but because the SOI file is based on a cross sectional design, there is no guarantee that a given observation in a particular year of the SOI data will be present in any succeeding year. Where possible, SOI data are used and missing years are filled in using data obtained from the IRS Masterfile, a collection of all individual tax returns filed; the Masterfile returns are not subjected to the same editing procedures as the SOI data.

Table 1: Percent distribution of final outcome codes for area-probability and list samples; 2001 SCF.

	AP	LS
OUT OF SCOPE		
Not a housing unit	3.9	NA
Vacant housing unit	7.2	NA
Seasonal vacant	2.4	NA
Sample incorrect	0.5	NA
Deceased	0.1	0.4
No eligible R in household	0.1	0.0
Permanently out of the country	0.0	0.2
Other out of scope	0.0	0.0
COMPLETE		
Complete interview, telephone	13.4	13.5
Complete interview, in-person	38.5	13.6
Complete interview, phone conversion	1.7	1.8
Complete interview, in-Person conversion	4.7	1.0
Partially completed interview	0.1	0.0
IN SCOPE NONINTERVIEWS		
Postcard refusal	NA	12.9
Final refusal, conversion attempted	7.5	8.9
Final break-off of interview	0.1	0.1
Final refusal by gatekeeper	0.0	0.1
Final unlocatable	0.2	0.4
R unavailable for field period	0.2	0.8
Language barrier (other than Spanish)	0.6	0.2
Physically or mentally incapacitated	0.2	0.2
Other noninterview	0.3	0.9
CENSORED		
Stopped work at end of field period	18.4	43.0
RESPONSE RATE (Complete/ (Total-Out of score))	68.1	30.7
N	4,993	5,026

whether the sample elements are even eligible. Thus, among high-quality survey organizations, great effort is usually expended to ensure that the status of nearly every case is known. In the case of the 2001 SCF (table 1), only a small fraction of a percent remained in this group. Where characteristics of a unit reflect a change in the framework underlying the sample, a set of rules may be specified in advance about actions to take. Typically, ineligible units are discarded and newly discovered units are sampled in a systematic way, so that the actual sample remains appropriately representative of the creation and destruction of housing units. In the 2001 SCF, a substantial fraction of the original area-probability cases was later determined to be ineligible; very few list sample cases were ineligible.[10]

[10]For the 2004 SCF, US Postal Service address sequences were used for listing most of the addresses in the area-probability sample (see O'Muircheartaigh et al. [2002]). Even this more up-to-date set of files yielded an out of scope rate of over 18 percent–4 percentage points higher than the case under more labor intensive listing system used for the 2001 survey.

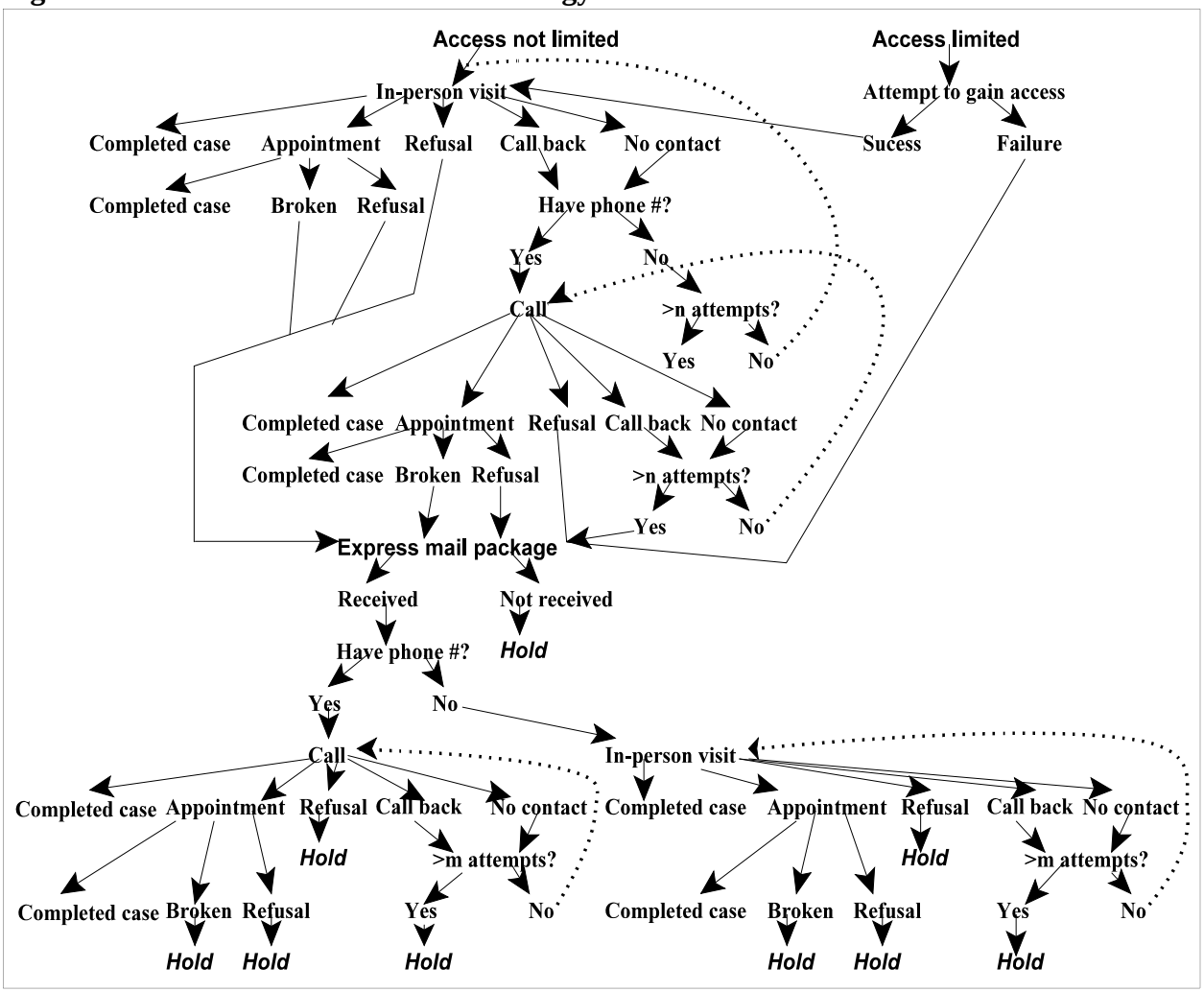
More complex sample problems are the potentially interdependent decisions made by the respondents and the survey field staff .[11] Clearly, some respondents would not participate in a survey under any imaginable feasible circumstance. But short of such an extreme position, it is not clear what is a "permanent" refusal to participate. As is typically seen when survey takers perceive response rates as being "too low," more intense application of effort may "convert" an initial refusal. Thus, it may be more useful to characterize respondents as having a range of possible responses conditional on personal characteristics and preferences, such as the shadow value of time or the sensitivity to privacy, and the inputs they receive–printed and virtual informational materials and interactions with interviewers and other field staff.

It is usually the case that interviewers and other field staff are under great pressure to produce complete cases. For this reason, it is reasonable to expect that at any given point in the field period, they will tend to apply effort first to cases that they believe are the most likely ones to be completed. If their expectations are unbiased, their actions will tend to exacerbate the patterns of nonresponse that would result from the predisposition of respondents. If resistant respondents and more willing respondents differ in important way in terms of the variables of interest, the behavior of the field staff will tend to amplify bias in the resulting data. In addition, because effort is endogenous, it becomes very difficult, without strong model assumptions, to draw conclusions about the characteristics of respondents that contribute to higher nonresponse rates. Detailed analysis of the "call records," or attempt-specific records maintained for every SCF observation, indicates strongly that differential application of effort is a serious problem.

One way to break the endogenous connection between effort and nonresponse is to pre-specify a program of field effort. Starting with the 2004 SCF, a phased plan was used to specify a more uniform program of treatment across all cases, at least through two levels of the plan. Figure 5 provides a schematic diagram of the process. In the first phase, interviewers were to approach respondents a specified number of times; if that effort failed to yield a completed interview, the respondent was sent a specially produced package of materials via express mail. After that mailing, another round of attempts was made. If the case failed to be completed

[11]Kennickell [2004] presents a model of field effort and presents the design of phased effort summarized in this paper.

Figure 5: SCF 2004 Phase I Contact Strategy.



before the end of this second phase, it was put aside for re-evaluation. Ultimately, cases in the third phase were targeted based on traditional techniques focused on feasibility of completion.

This phased system of effort does not entirely break the endogeneity of effort and completion, but it does partition the sample into segments that can be analyzed as if effort were exogenous. Models can be constructed that allow one to address systematic components of nonresponse that are related to respondents' behavior, if data for all sample observations are available from another source.[12] In the case of the SCF, external data are available for the area-

[12]Even in the absence of supplementary information, gains might be made from comparison of distributions of variables across the three segments.

probability sample at the level of Census tracts from the 2001 Census of Population and for the list sample at the individual taxpayer level from the frame data used in the selection of the sample. In addition, as a part of the overall survey design, interviewers are required to record observations for each case about the structure at the sample address and the initial informant.

In addition to maintaining a more measurable level of effort, such sample management has an ethical benefit. Normally researchers worry about the rights of people to give "informed consent." But when the data collected serve an important function in policy analysis and longer-term analysis that underlies the construction of future policies that affect all types of people, there is an ethical imperative to ensure that refusals are also informed. A phased effort of the sort applied for the SCF is credible in giving all respondents selected an equal opportunity to understand the reasons for participation.

V. Post-Survey Adjustments

When data collection ends, what is available is information from (usually) a subsample of the original respondents, a track of effort applied to persuade people to participate, and any auxiliary data that may have been built into the design or that are available through linkage to the sample. Taken together, sample selection, sample management and response processes–summarized here as Ŝ–applied to the original sample yield a *realized sample* that may differ from the universe of eligible subjects of the survey in more subtle ways than simply the smaller number of observations. If all observations participate fully, sufficient information is available in the initial probability design for the analysis of the data. In the more usual situation where cooperation is not complete, to make the observed data analytically useful, a way must be found to make the data provided by the participating units approximate the information that might have been provided by the full population. This step usually entails imputation for item nonresponse and weight adjustments for unit nonresponse. The focus in this section is on adjustments for unit nonresponse, but one may apply sample control arguments to item nonresponse as well.

Weights are a subset of a class of model-based estimators that may be used to map observed information into an estimate for the relevant population universe, conditional on design

information and the behaviors of the actors in the data collection process. Simply put, weights provide a relative measure of size of each survey observation in computing estimates reflective of something beyond the cases in the realized sample; most often weights are calculated to sum to the total number of units in the relevant universe population. One possibility is to find a weighting function ω(Z) to solve to the statistical problem of minimizing the sampling variance of a key survey estimator ζ subject to that estimator being unbiased (where the expectations are taken over the population universe under the observation mechanism Ŝ):

$$\underset{\omega(Z)}{Min}\,\mathrm{Var}_U\left[\zeta\left(Y,\omega(Z),\hat{S}\right)\right] \quad s.t. \quad \mathrm{E}_U\left[\zeta\left(Y,\omega(Z),\hat{S}\right)-\zeta\left(Y\right)\right]=0$$

If an estimator is unbiased under a given weighting design but has high variability under Ŝ, one faces a relatively large probability that weighted estimates based on a realized sample might be far from the unbiased center, with no means of knowing that to be the case. An alternative formal possibility is to allow for a trade-off of variance and bias via a function Ω increasing in both of its elements:

$$\underset{\omega(Z)}{Min}\,\Omega\left\{\mathrm{Var}_U\left[\zeta\left(Y,\omega(Z),\hat{S}\right)\right],\left|\mathrm{E}_U\left[\zeta\left(Y,\omega(Z),\hat{S}\right)-\zeta(Y)\right]\right|\right\}$$

Although one would never willingly add bias in the absence of other constraints, it is sometimes worth considering approaches that entail some bias but that provide a lower degree of possible deviation from the expected center of the distribution of the true value of ζ. Such a tradeoff may become particularly valuable to consider when a survey is expected to be part of a series of surveys where comparisons across waves is important and stability is, consequently, very important.

In practice, it is unlikely that one would have either the appropriate known structure for full adjustment or the necessary data to implement such a structure. But continuing study of the factors in Ŝ may lead to deeper understanding. Such research has been at the heart of the SCF since its beginning and it has led to many changes in weighting procedures–and field management techniques, as noted above.

As indicated in table 2, there is substantial variability in nonresponse across the two SCF samples and within each sample–as well as in all of these groups over time. For the area-

Table 2: Response rates 1992-2001 SCF; by sample type and subgroups within sample type; percent.

	1992	1995	1998	2001
Area-prob. sample				
Area				
Northeast region	65.4	60.1	62.4	68.7
Northcentral region	68.5	70.9	67.4	66.9
Southern region	70.3	67.2	68.3	70.7
Western region	66.4	65.3	63.8	64.9
Largest MSAs	61.8	58.9	62.3	63.2
Other MSAs	67.4	66.6	66.6	69.7
Non-MSAs	75.7	77.6	70.3	73.3
All areas	68.0	66.3	65.9	68.1
List sample				
Stratum				
1	42.8	45.3	41.3	37.3
2	41.4	39.5	39.2	40.9
3	37.4	35.5	36.2	36.9
4	34.7	35.0	35.8	36.2
5	31.4	30.4	30.4	31.9
6	26.0	23.9	23.9	24.2
7	14.4	12.8	8.3	10.0
All strata	31.3	30.4	28.6	29.6

probability sample, the most noteworthy fact in the table is the decline in response rates with the population size of the sample area. Typically, cities such as New York and Los Angeles have much lower response rates than rural counties. In the list sample, response rates decline strongly with wealth strata. In both samples, these are surface effects reflecting a much more complicated underlying structure relating to privacy concerns, time pressures, and many other factors related to respondents, interviewers, and other actors in the data collection process.

Although deep modeling of these factors is not feasible, repeated experimentation and testing have revealed dimensions of nonresponse that appear most important. The approach taken to weighting in the SCF employs largely techniques of post-stratification and raking.[13] Under post-stratification, observations grouped according to classes of characteristics have their weights adjusted by a uniform proportion to bring the weighted sample estimate of the number of units in each group into approximate line with more reliable externally available totals. Raking takes this approach

[13]See Little [1993 for a discussion of post-stratification and raking. For a detailed presentation of the SCF weighting methodology, see Kennickell and Woodburn [1997] and Kennickell [1999b].

sequentially and iteratively through multiple dimensions until sufficient convergence is reached. If the adjustment cells are a sufficient proxy for unobserved (or imperfectly observed) qualities that drive nonresponse, this approach reduces bias. These methods may also help in reducing sampling error or otherwise increasing estimation efficiency by constraining the sample estimates to reproduce key observed dimensions of the population. For technical reasons, the adjustment cells need to be limited to only the most important; among other things, a hidden cost of excessive post-stratification or raking is an inflation of sampling variance.

The key to discovering the relevant structure and to applying the resulting weighting adjustments is the existence of auxiliary data. Here the sample design can play an important part by taking into account the potential for meshing the survey data with other data sources. In the SCF, care is taken to select the area-probability sample in a way that allows some linkages with data from the Decennial Census as well as to a more limited degree the Current Population Survey. For the list sample, connection to the original SOI data is more direct, and the sample design builds in groupings, such as the original wealth index strata, that are believed to be important in nonresponse. A variety of supplemental data are also collected for all SCF sample cases during the period of the main survey data collection. Together, these pieces of information support a standardized program of weighting adjustments that is comparable from 1989 to 2004 and they provide a basis for continuing investigation of the effectiveness of the adjustments and the possible alternatives.

To facilitate the construction of confidence intervals for the survey estimates, the SCF provides a set of replicate weights in addition to the main weights. The SCF sample design is sufficiently complex that variance estimates are not easily made using either analytical approximations or common numerical techniques. The replicate weights are based on 999 samples drawn from the realized sample in such a way as to proxy for what are believed to be the most important dimensions of selection variability. The weights for each pseudo-sample are adjusted using the full set of procedures applied to the main weights. Simulation of the variability of any survey estimator requires only a straightforward replication of the calculation for each weight replicate.

VI. Conclusions and Future Work

The discussion of sampling typically ends with the specification of the design. But the implementation of the design and accounting for the design at the analysis stage raise deep questions that inevitably call for an expanded view of what is encompassed by sampling. Survey sampling is intended to provide a mechanism for observing vectors of characteristics within a framework that allows a probability of selection to be assigned to each case. But what is relevant for analysis is the probability of *observation*, of which the probability of *selection* is only one part. The probability of observation will also be a function of the likelihood that a case can be identified and located, the probability that effort will be applied to the case, and the probability that the respondent will agree to participate to some degree.

This paper outlines the way that the sampling plan for the SCF extends throughout all phases of the survey. In surveys like the SCF that set out to provide useful information on the full range of wealth, it is important to think particularly carefully about how to obtain sufficient information to design a sample that breaks up the target population into groups roughly classifiable by an indicator of wealth. The SCF uses a dual-frame design. The most distinguishing feature of the design is its use of a mapping from income to wealth for stratification in a part of the sample selected from statistical records derived from tax returns. This structure allows not only the differential selection of wealthy households, but it also allows naturally for a nonresponse adjustment at the end of data collection to account for differentially lower participation rates among such households. In executing the SCF, or virtually any other survey, the attention field staff apply to the sample observations to convince them to participate is a far from uniform force. If, as seems reasonable, the decision of respondents to participate is a function of inputs of efforts to persuade them, then variations in the application of effort have the same effect as altering selection probabilities. But often this stage of "selection" is treated as being entirely neutral and it is rarely addressed directly in post-survey adjustments. As a result of earlier experience, the SCF has imposed a more uniform structure on the management of the individual survey cases to ensure more uniform and measurable effort. The nonresponse adjustments at the end of the survey provide a chance to make amends for a variety of errors, if an approximately correct mapping of the achieved sample into the population universe can be

achieved. The SCF employs a series of adjustments that are made possible by a design that builds in sufficient conceptual overlap with other data.

Progress in meaningful wealth estimation requires much of the same developments required for other surveys, but because of the greater than average difficulty of gaining respondents' cooperation in wealth surveys and other such factors, the pressure to improve is generally greater. Instrument design is a perennial problem, even in the long-established SCF, because the nature of the financial world continues to evolve in ways that are increasingly difficult to specify in simple comprehensible terms. On the more statistical side of wealth surveys, I see a clearer path, but much work remaining along the way. Although there may be particular and very difficult problems in setting an initial sample design for wealth measurement, by now the experience of the U.S. SCF, the Cyprus SCF (Karagrigoriou [2004]) , and the Spanish EFF (Bover [2004]) give much information and intuition on different ways to proceed. In my belief, the places where effort is most needed still is in understanding the management of cases during the field period and in understanding and coping with the nonresponse that inevitably occurs. A key for both of these problems is the collection of additional data or redesigning existing systems to provide more data in an analytically more useful form. Of particular importance are attempt-level data on survey administration, information about interviewers, and additional information about characteristics of sample units (regardless of whether or not they were interviewed). Such information would allow the building of more complex behavioral models of nonresponse, a task that should hold interest for economists interested in collecting wealth data. The payoff of this research will be greater integration of survey design, execution and analysis and consequently better measurement.

Bibliography

Aizcorbe, A., Arthur B. Kennickell and Kevin B. Moore [2003] "Recent Changes in U.S. Family Finances: Evidence from the 1998 and 2001 Survey of Consumer Finances," *Federal Reserve Bulletin*, pp. 1-32.

Bover, Olympia [2004] "The Spanish Survey of Household Finances (EFF): Description and Methods of the 2002 Wave," Documentos Ocasionales no. 0409, Banco de España.

Internal Revenue Service [2001] "Individual Income Tax Returns, 1998," Internal Revenue Service, Statistics of Income Division, Washington, DC, Publication 1304 (Rev. 04-2001).

Karagrigoriou, Alexandros [2004] "Sampling and surveying: Cyprus," working paper University of Cyprus.

Kennickell, Arthur B, [1999a] "Using Income to Predict Wealth," http://www.federalreserve.gov/pubs/oss/oss2/method.html.

_____ [1999b] "Revisions to the SCF Weighting Methodology: Accounting for Race/Ethnicity and Homeownership," http://www.federalreserve.gov/pubs/oss/oss2/method.html.

_____ [2000] "Wealth Measurement in the Survey of Consumer Finances: Methodology and Directions for Future Research," http://www.federalreserve.gov/pubs/oss/oss2/method.html.

_____ [2001] Modeling Wealth with Multiple Observations of Income: Redesign of the Sample for the 2001 Survey of Consumer Finances, http://www.federalreserve.gov/pubs/oss/oss2/method.html

_____ [2004]Action at a Distance: Interviewer Effort and Nonresponse in the SCF, http://www.federalreserve.gov/pubs/oss/oss2/method.html.

Kennickell, Arthur B. and Douglas A. McManus [1993] "Sampling for Household Financial Characteristics Using Frame Information on Past Income," http://www.federalreserve.gov/pubs/oss/oss2/method.html.

_____ and R. Louise Woodburn [1997] "Consistent Weight Design for the 1989, 1992, and 1995 SCFs, and the Distribution of Wealth," http://www.federalreserve.gov/pubs/oss/oss2/method.html.

Kish, Leslie [1965] *Survey Sampling*, John Wiley and Sons, New York.

Little, Roderick A.J. [1993] "Post-Stratification, a modeler's perspective," *Journal of the American Statistical Association*, v. 88 (September), pp. 1001-1012.

Neyman, J. [1934] "On the two different aspects of the representative method," *Journal of the Royal Statistical Society*, vol. 97, pp. 558-625.

O'Muircheartaigh, Colm, Stephanie Eckman, and Charlene Weiss [2002] "Traditional and Enhanced Field Listing for Probability Sampling," *Proceedings of the American Statistical Association Social Statistics Section*, pp. 2563-7.

Särndal, Carl-Erik, Bengt Swensson and Jan Wretman [1992] *Model Assisted Survey Sampling*, Springer-Verlag, New York.