Finance and Economics Discussion Series
Divisions of Research & Statistics and Monetary Affairs
Federal Reserve Board, Washington, D.C.

Updates to the Sampling of Wealthy Families in the Survey of Consumer Finances

Jesse Bricker, Alice Henriques, and Kevin Moore

2017-114

Please cite this paper as:
Bricker, Jesse, Alice Henriques, and Kevin Moore (2017). "Updates to the Sampling of Wealthy Families in the Survey of Consumer Finances," Finance and Economics Discussion Series 2017-114. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2017.114.

Updates to the Sampling of Wealthy Families in the Survey of Consumer Finances*

Jesse Bricker, Alice Henriques, and Kevin Moore

Federal Reserve Board

Abstract

Participation in household surveys has fallen over time, making it harder to produce a household survey—like the Survey of Consumer Finances (SCF)—in a timely manner. To address these challenges, the reference year of the sampling frame data for the 2016 SCF wealthy oversample was shifted back one year, allowing the oversample to be selected earlier than the past.

In implementing this change, though, we risk identifying an outdated set of families and introducing variability in the sampling process. However, we show that the set of families selected in the new frame are observationally equivalent to those that would have been selected from a past frame, and that the increased variability of wealth estimates is compensated-for with the use of more comprehensive data than in the past.

Other aspects of the SCF sampling process are revisited, too. We continue to find support for using permanent income in the sampling process, rather than annual income. We also estimate the geographic distribution of wealthy families and show that the current distribution is similar to the past. We propose adding one geographic area to the oversample, though, and supplementing by 100 the set of sampled families.

JEL codes: D3, H0

Keywords: household surveys, wealth, distribution, sampling

* The analysis and conclusions set forth are those of the authors and do not indicate concurrence by other members of the research staff or the Board of Governors. We would like to thank our colleagues on the SCF project who helped make this research possible: Lisa Dettling, Joanne Hsu, Lindsay Jacobs, Sarah Pack, Jeff Thompson, and Richard Windle. We have also benefited from discussions with Steven Pedlow, John Sabelhaus, Tom Crossley, and from the inherited knowledge of Arthur Kennickell. Finally, this work would not be possible if not for the generous cooperation from our partners at the Statistics of Income at the Internal Revenue Service, especially Barry Johnson, David Paris, Michael Parisi, Lori Hentz, and Lori Russ.

I. Introduction

Capturing the full distribution of household wealth in the United States is difficult. The concentrated nature of wealth means that a random sample of U.S. families is unlikely to capture the small minority of families that hold the large majority of wealth. Wealthy families are also less likely to participate in a survey, meaning that a random sample of families may incorrectly estimate the mean and variance of wealth. The Survey of Consumer Finances (SCF) provides unbiased and efficient estimates of the distribution of household wealth by using a dual-frame sample design, whereby a nationally representative set of families (the "AP sample")—selected from an address-based frame—is supplemented with an oversample of wealthy families selected from administrative records derived from income tax returns (the "list sample").

Participation in household surveys has fallen over time, making it harder to produce a household survey in a timely manner. This reluctance stems from a variety of sources: declining use of landline telephones makes families harder to reach, and well-publicized data breaches make families less likely to reveal sensitive information. Complicating matters for the SCF is the availability of the list sample frame, which is often available only days prior to the start of the field period and has led to delays in contacting the sampled list families.[1]

In an effort to complete the SCF field period earlier, the reference year of the frame data for the 2016 SCF oversample was shifted back one year. In earlier SCF surveys, the list sample frame data were based on income earned *two* years prior to the SCF survey year.[2] The frame for the 2016 SCF list sample, though, is based on income from *three* years prior to the survey year. This paper considers the costs and benefits to such a change.[3]

SCF data quality may be negatively impacted by this change, as the quality of SCF data depends, in part, on accurately identifying wealthy families in the list sample frame data. This is

[1] The address-based frame for the AP sample is available well before the SCF field period begins.
[2] For example, the 2013 SCF list sample frame were based on income earned in 2011. Income earned in tax year 2011 is filed in calendar year 2012, and often not until late in the year. Thus, the frame data aren't processed until early 2013.
[3] Historically, the SCF field period lasts many months past its planned close date, and the last interview is often conducted nearly a full year after the first. Moving the reference year back can allow the list sample to be drawn well ahead of the field period. List sample locating can begin as the field period starts, possibly leading to an earlier finish to the SCF field period. These benefits, of course, may be offset by countervailing forces (increased reluctance of families to participate, increased prevalence of gated residences, among others).

no small task, as the frame data describe taxable *income*, while the goal of the oversample is to identify (and oversample) *wealthy* families.[4]

However, we show that we can move the reference year of the frame back with little to no cost in the data quality of the sampling frame. Throughout the paper we use either the 2013 SCF list sample frame—which is drawn from a frame of tax year 2011 income (along with 2010 and 2009 income for most families in the 2011 frame)—or a hypothetical list sample frame based on income from 2010 (along with 2011, 2009, and 2008 income for families in the 2010 frame).

First, we use the two datasets to show that the composition of families sampled for the 2013 SCF—based on the 2011 income frame—are observationally equivalent to those that would have been sampled from the hypothetical 2010 income frame. That is, the mean income in 2011-2009 of the families in the hypothetical 2010 frame are equivalent to the mean 2011-2009 income, by sampling strata, of the families that were actually sampled for the 2013 SCF. Partly this reflects the strong overlap between adjoining years in the sampling frames, as nearly all upper-strata families from one frame data file are found in the frame data in an adjoining year.

Second, we show that, all else being equal, the variability of wealth predictions will increase when moving back the sampling reference year. About 10 to 15 percent of families are predicted to be in a different wealth sampling strata—and so have different probability of selection—when the frame reference year is moved back by a year. However, in moving back the frame reference year we will be able to use a *full* panel of income to sample—something not available in the past few surveys—which will offset most of the increase in variability.[5] In our data, about 10 percent of families change sampling strata when we use a full income panel in sampling, which nearly offsets the increase in variability described above.

Ultimately, though, we will also use the most recent frame data—those that would have been used had we not shifted the sampling frame back—to produce the 2016 SCF. These data will be released too late to be used in sampling, but we can use these data after the fact to update and re-

[4] As described later, SCF sampling practice uses two models to predict a *wealth* ranking from these *income* data.
[5] For the past several surveys, there has been a glitch in the matching algorithm between INSOLE data and unedited IRS administrative data, which was discovered during this work. To be clear, basing the sampling frame on a lagged year does not lead to a higher match rate.

allocate families across sampling strata for weighting purposes, and for statistical editing of the SCF data. We expect that few cases will be re-stratified, though.

Aside from this change in reference year of the list sample frame, the paper includes a re-assessment of several aspects of the SCF sampling process. First, we revisit the use of permanent income—instead of annual income—to model wealth and find continued support for using permanent income (Kennickell 2001). When wealth is predicted with annual income, transitory income changes can lead to volatility across years in the wealth distribution.

Second, we revisit how the dual sampling frames of the SCF—the list sample and AP sample—fit together.[6] The sampling practice for the SCF has always restricted the list sample to have the same geographic coverage as the national address-based sample. However, if the geographic distribution of wealthy families does not coincide with the geography of the national sample, then we could expect a bias in SCF wealth estimates.

In the recent sampling frame data, wealthy families cluster more tightly geographically than does the overall US population but, as in Frankel and Kennickell (1995), the majority of wealthy families live in areas sampled for the national sample. The SCF has traditionally used a geographic post-stratification weighting system developed by Kennickell and Woodburn (1999) to smooth out any inconsistencies between the sampled areas and the overall U.S. population; our results suggest that this technique remains effective.

However, we propose supplementing the list sample by one geographic area. As the sampled areas of the SCF are not publicly released, the exact location is suppressed in this public version of this paper. It is located near other sampled areas so we expect little added labor cost in adding this area because local field staff are already available.

Finally, we also revisit the number of families sampled, and propose adding more than 100 cases to the top-end sampling strata, taking into consideration both increased mortality probability (due to shifting the sampling data one year back) and declining rates of interview completion within the list sampling strata.

[6] Frankel and Kennickell (1995) provide the first such assessment. Our re-examination comes in light of higher wealth concentration estimates from recent surveys (Bricker et al, 2016) and from wealth predicted from a set of income tax data (similar to our frame data—Saez and Zucman, 2016).

A voluminous literature supports all aspects of the SCF sample design. Our paper fits into this literature in several ways. First, by focusing on how modeled wealth estimates vary over time, we continue past work on using income to model wealth for SCF sampling purposes (as in Kennickell and MacManus, 1993; Kennickell, 2001; Kennickell, 2007). Assessing the fit of the AP and list samples continues the work of Frankel and Kennickell (1995), and our consideration of misclassified filers touches on the SCF weighting design (Kennickell and Woodburn, 1999).

Section II describes the SCF sampling and weighting procedures in the SCF, section III considers the question of moving the sampling year back, section IV considers adding sample size to the list sample, section V describes the geographic overlap between the list sampling data and the address-based national sample, and section VI concludes.

II. The SCF Sampling Procedure

The SCF is a cross-section survey, conducted every three years by NORC at the University of Chicago on behalf of the Federal Reserve Board (FRB) and with the cooperation of the Statistics of Income (SOI) at the Internal Revenue Service (IRS). SCF families respond to questions about financial and nonfinancial assets, debts, employment, income, and household demographics. The SCF provides the most comprehensive and highest quality microdata available on U.S. household wealth.[7]

Economic resources in the U.S. and other industrialized countries are highly concentrated. Measuring and explaining income and wealth concentration has challenged economists at least since Pareto (1896) and Kuznets (1953). Measuring income and wealth using simple random sampling and household surveys is not a viable solution, because thin tails at the top lead to large sampling variability, and disproportional non-participation at the top biases down top share estimates. The Survey of Consumer Finances (SCF) overcomes both problems by oversampling at the top using administrative data derived from income tax records, and by verifying that the

[7] See Bricker, et al (2017) for results from the most recent triennial SCF. A great degree of security is involved with this sampling procedure and formal contract govern the agreement between the FRB, NORC and SOI. The FRB selects the sample from an anonymized data file. The FRB sends the sampled list to SOI, who remove the famous families and passes along the list to NORC for contacting. NORC collects the survey information and sends to FRB. Thus, the FRB never knows any contacting information, SOI never knows any survey responses, and NORC never knows anything more than survey responses and location information.

top is represented using targeted response rates in several high end strata.[8] The list sample ensures that the SCF has adequate representation of the upper tail of the wealth distribution and adequate representation of sparsely held assets.

Dual frame design: AP sample and List sample

The SCF combines a geographically stratified and nationally representative area probability ("AP") sample with a list sample, an oversample of households that are likely to be wealthy. The AP sample is drawn by NORC at the University of Chicago and provides a nationally-representative sample of families.[9]

The only official wealth record that exists in the U.S. comes from an estate tax applied at death, so there is no administrative data system directly associated with measuring the cross-section of wealth at a point in time. Thus, the list sample depends on inferring wealth from administrative records derived from income tax returns—the Individual and Sole Proprietor (INSOLE) data file maintained by SOI (Statistics of Income, 2012).[10] The INSOLE file is a sample of tax filings from the IRS administrative tax data, statistically edited for quality by SOI. In the list sampling process, wealth is inferred from these income records through two models that relate wealth to income (described in more detail later in this section).

In earlier SCF surveys, the list sample frame data were based on income tax returns filed the year prior to the SCF survey year, meaning that the income was earned *two* years prior to the SCF survey year. In the 2013 SCF, for example, the frame data were derived from tax returns

[8] The administrative data used in the sampling also show that SCF participants are observationally equivalent to non-participants within the high end strata (Bricker et al, 2016).
[9] See Tourangeau, et al. (1993), and O'Muircheartaigh et al. (2002) for more information on the NORC samples.
[10] The INSOLE file consists of a sample of individuals and sole proprietorship tax filings from the IRS administrative tax data, statistically edited for quality by SOI. Tax filings with unique income are oversampled and many high-income records in the INSOLE file are sampled with certainty (Statistics of Income, 2012). The INSOLE file is a sample of the IRS administrative tax data, so the LS is a sample from a sample. No correction for this is made during the LS sampling procedure, though, because the certainty sample and rare incomes found in the INSOLE file are a near certainty sample of the LS target population (Kennickell and Woodburn, 1999). The unit of observation in the INSOLE data is a tax unit while the SCF unit of observation is a family. In practice, there are millions more tax units than families because several members of a family can file distinct tax returns; without a correction, these multi-filer families would have a disproportionately large chance of being selected. To account for this in the SCF LS sampling process, the INSOLE sampling weight of tax units that filed "married filing separately" is divided in half. Further, all filers below the age of 18 are dropped (a family headed by someone less than age 18 is ineligible for the SCF). Still, to a certain extent, the discrepancy between tax units and families remains in the adjusted INSOLE sampling frame.

covering income earned in 2011.[11] From 2001 and forward, the SCF list sample is drawn using multiple years of income for the returns in the frame. In the 2013 SCF, for example, the 2011 income data in the frame were supplemented with 2010 and 2009 income.[12]

The INSOLE file is not designed to be a panel, though certainty sampling of high income families (among others) means that families with consistently high incomes are often in the sample year over year. Filers with total income of at least $5 million, filers with total income of less than negative $5 million, filers with $50 million of Schedule C receipts, and filers with at least $200,000 of AGI but zero tax liability are all sampled with certainty (Czajka, Sukasih, and Kirwan, 2014; Bryan, 2015). Filers with at least $2 million (or less than negative $2 million) in income are sampled at about a 50 percent rate. The file is also sampled in a Keyfitz method, meaning that there is a strong overlap between adjoining year files.

Though the nominal filing deadline for *year t* income is April of *year t+1*, many families receive an extension and submit a final filing in October of *year t+1*. The INSOLE file, then, is often not available for sampling until early in *year t+2*.[13]

Wealth models

The SCF sampling strategy uses two methods of predicting wealth from income. The process of selecting the list sample has evolved since the current SCF began in 1989, as more refined models for selecting wealthy respondents have been introduced, including moving from cross-

[11] Because the 2013 list sample frame covers returns filed in 2012, not all returns will refer to 2011 income—for example, an amended return from a previous year may be sampled into the frame.

[12] If a link to the past INSOLE file was not possible, though, a link was created to past data from the unedited IRS administrative tax data. Thus, the panel of income data used to sample the 2013 SCF LS was a mixture of edited INSOLE data and unedited tax data. About 55 percent of the final sampling data in the 2013 SCF are linked to a past INSOLE file income data while 45 percent are linked to unedited income tax data. About 22 percent of the sampling frame did not have two years of panel data – either through INSOLE data or unedited IRS administrative tax data. It was uncovered during the work for this paper that the matching algorithm between the INSOLE and unedited IRS administrative data files was misaligned in a non-random way. With that algorithm corrected, the 22 percent unmatched rate would have been lower.

[13] Because this schedule is known beforehand, much of the early sampling work is based on a preliminary and incomplete INSOLE file. In February 2013 the FRB received the 2010 and 2009 panel tax records, linked to the 2011 preliminary file. As such, most families in the upper strata are linked to unedited IRS administrative tax data, not the statistically edited INSOLE file. In the wealthiest sampling strata, 90 percent of 2010 and 2009 income came from the unedited income data. Using a mixture of INSOLE-unedited income panel data decreases sampling efficiency by increasing noise around the income classification and wealth rankings, especially among the wealthiest families (Kennickell, 2001).

section to panel-based administrative records in order to better control for transitory income fluctuations (Kennickell, 2005; Kennickell, 1998)

The first method of predicting wealth from income is a "gross-capitalization model," generated by inflating asset-based income in each tax record by an asset-specific rate of return, and then by adding a predicted housing value (Greenwood, 1983). The general form of the SCF model is:

$$\widehat{wealth}_i^{GC} = \widehat{house}_\iota + \sum_{\forall k}[\overline{Income}_i^k / r^k],$$

where there are i=1…N tax units, K types of income and $r_k$is the rate of return on the k-th type of income, and r^k is typically $\epsilon(0,1)$. There are six types of income in the SCF model: taxable interest, non-taxable interest, dividend income, rents and royalties (in absolute value), business, farm, and estate income (in absolute value), and capital gains (in absolute value).[14] The income fed into the model is a weighted average of three years of income: for each type of income *k* and household *i*: $\overline{income}_i^k = income_i^{k,T} * \left(\frac{1}{2}\right) + income_i^{k,T-1} * \left(\frac{3}{10}\right) + income_i^{k,T-2} * \left(\frac{2}{10}\right)$. Using multiple years of income data to identify wealthy individuals helps to smooth over the effects of transitory income fluctuations that are especially prevalent for capital incomes and at the top of the distribution.

The second method of predicting wealth from income uses the empirical correlation between wealth collected in the SCF and income from the administrative sampling data. The basis for this "empirical correlation model" is a regression of observed SCF wealth from the most recent SCF on the administrative income used to generate the SCF list sample for that survey year. The most recent SCF is denoted here as T-3 and the base sampling income data are from two years prior to that:

$$\ln(SCF\ wealth_i^{T-3}) = \ln(\overline{Income}_i^{T-5})\beta + \varepsilon_i.$$

The matrix of sampling income for the previous SCF ($\overline{Income}_i^{T-5}$) consists of more than 30 logged income variables and a dummy indicating the presence of such income for that tax unit, plus some basic demographic data.[15] The $\hat{\beta}$vector from this regression model is then applied to the current administrative sampling data to obtain a predicted wealth index:

[14] Model details are provided in Appendix A, including rates of return. Income is a weighted average of three years of sampling income.

[15] As in the gross capitalization model, income is a weighted average of three years of sampling income. The variables in the empirical correlation model are selected by a stepwise model selection method; complete details are provided in Appendix A.

$$\widehat{wealth_i^{ECorr}} = f(\overline{Income_i}\ ;\hat{\beta}).$$

In contrast to the gross-capitalization model, one key difference is that the empirical correlation model allows a variety of income variables that are not necessarily based on a physical asset and allows rates of return to vary across different types of families.

The gross capitalization and empirical correlation models generate two independent sets of rankings—or wealth "indices"—so that each tax record has two rankings. Ultimately the information from both is used to generate an overall wealth ranking. The two normalized indices are then blended together and the sampling data are ordered from least wealthy to most wealthy by the blended index.[16]

Once ordered by wealth, the records are organized into seven sampling strata. The top 500 families are placed in the top (seventh) strata, and the remaining are organized into the other six wealth strata. The strata are organized by increasing expected wealth, and the top sampling strata (the 4th, 5th, 6th, and 7th) cover the top one percent. Records in the lowest stratum are often comparable to the AP sample. The probability of being sampled increases as the value of the strata increases.[17]

For labor cost reasons, the list sample is sampled from the same geographic areas as the AP sample. Thus, after each tax record is ranked by expected wealth, the records from areas outside of the AP sample are removed. In doing so, about one-third of the tax record sample is lost. To compensate, the weights of the remaining tax records are inflated by the inverse of the probability that the unit's area was selected for the AP sample.

But doing so has potential costs: the AP sample is selected to be a nationally representative sample based on demographic characteristics of geographic areas. Adjusting the sampling weights from the tax data by a geographic selection probability may be problematic, as prior

[16] Before being blended together, each index is normalized by subtracting its median and dividing by its interquartile range. Typically, the blend is a 50/50 split, although in recent years the split has favored the correlation model, due to the strengths that we will observe later.

[17] In practice, the number of observations in this certainty strata is higher than 500 as some observations are cannot be interviewed because they (a) responded or refused to the most recent SCF, (b) responded to the second-most-recent SCF, or (c) are outside of an NORC-sampled NFA.

work has shown that the intersection of these two selection models does not have to overlap.[18] The alignment of the list sample and the AP taken up in detail in the section V.

The SCF list sample is selected through a probability proportional to size (PPS) method, stratifying by the seven wealth strata, nine financial income sub-strata and four age sub-strata.[19] Initially, the size measure is the sample weight of the tax record in the INSOLE file, inflated by the record's probability of selection into the NORC national sample (described above).[20]

A target number of cases are selected within each wealth strata—see column (2) of table 9 for the exact amounts. In total, about 5,100 LS cases are selected and the majority are from strata that capture the top 1 percent of expected wealth. The sample is sent to SOI, who scrub the sample of any case in the *Forbes* 400 or any case deemed too unique for a public SCF data release. The remaining cases are sent to NORC for contacting.[21]

Sample Weights

The SCF AP and list families are woven together by a set of sample weights. In each sample, the base sampling weight is adjusted to account for non-response, population targets, and the strengths of each sample. The base weight for list sample families is the measure of size used in the sampling process.[22] An initial adjustment is made to account for seeming misclassifications, which are inherent in the sampling process because of the imperfect modeling of wealth from income,[23] and are further post-stratified to match population totals across the sampling strata

[18] This topic is discussed in more detail in section V. The INSOLE sample that remains after discarding the non-AP-sampled areas is likely to over-select wealthy cases in smaller areas (relative to the INSOLE-drawn sample). After the survey field period is complete and final SCF weights are being drawn, the list sample weights are post-stratified to Census region, which could alleviate some of this bias.

[19] Sub-strata are arranged (head-to-tail) so that the PPS mechanism selects a good number of cases for each financial income and age bin.

[20] After being inflated, though, some records may have large measures of size. To ensure that units are not sampled with certainty, units are ranked by size within each wealth strata and divided into terciles; each unit in a tercile gets the mean measure of size within that tercile.

[21] Contacting these cases in the field is difficult, as the list cases often have layers of gate-keepers to ensure their privacy. The SCF has a three-phase field process to ensure that each case is worked sufficiently and that effort is directed to all cases, not just cases that are easy to gain cooperation (Kennickell, 2005**)**. Even so, within stratum response rates at the top end ranged from 33 percent to about 12 percent in the 2013 SCF. These response rates are somewhat targeted and the response rates would surely be lower in the absence of exceptional effort by the NORC field staff. The SCF sample weights adjust for nonresponse and thus account for the variability in response across list sampling strata.

[22] Namely, the INSOLE weight, divided by probability of selection, and grouped into terciles, as mentioned above

[23] After the SCF data are collected, the net worth of some families in a given strata are more similar to adjoining strata than their own strata. The sampling strata of these families appears to be misclassified; as such, these families are given the median weight of the families in the adjoining strata. Specifically, families with net worth greater than

(defined by predicted wealth), by Census region, and by financial income groups in the sampling data[24], and then raked (three times, sequentially) to balance the total weight across financial income groups, wealth strata groups, and Census regions.[25] Then, the population in the frame for strata three and above is adjusted to the *survey year* total population by growing the list sample frame population by the growth rate of the Current Population Survey (CPS) between the sample frame year and the SCF survey year.

Finally, the list and AP samples are fit together. The AP sample provides better coverage of the entire set of US families (i.e. irrespective of whether they filed taxes) while LS families provide better coverage of families at the upper tail of the income and wealth distribution. When the AP and list families are blended together, the AP families that did not file a tax return are set aside, as there is no overlap between the two samples for this group.

The remaining AP and list families are ordered by gross assets with families organized into seven classes, defined by total gross assets. The weights of each list (AP) family is adjusted by the share attributable to list (AP) families within each gross asset class. Then the weights of classes three through seven are adjusted by the total number of families implied by the list sample.

The weights of classes one and two are adjusted by the total of households in the CPS, less the number of households in classes three through seven, less the number of non-filers. Further, the weights of families that filed a tax return are then post-stratified by fine age categories in the CPS, the weights are raked to CPS homeowner-by-rough age categories and region, and post-stratified to fine age categories again.

Overall, the sampling and weighting mechanisms serve to select a set of high-wealth tax filers and fit those families into a set of US families.

---

the 90th percentile of their own strata are given the median weight of the next highest strata and families with net worth lower than the 10th percentile of their own strata are given the median weight of the next lowest strata. Between 30 and 50 families are reclassified each survey year.

[24] In general, post-stratifying weights means scaling up the weights of the respondents to match that of the total frame population. Doing so assumes that the respondents and non-respondents are interchangeable; it would be a problem, for example, if the respondents in a given strata all had lower income than the non-respondents. Bricker et al (2016) suggests this is not the case.

[25] There is no optimal number of times to rake and we choose three. Typically, the more one rakes, the more likely one is to end up with cases with extreme weights.

III. Change in Sampling Year

An emerging challenge of running the SCF is the increasing difficulty in completing the SCF by the end of the survey field period. Because the most recent list sample frame data are only available as the field period begins, there is a delay in initially contacting these list sample families. Beginning in the 2016 SCF, the year of the list sampling frame will be shifted one year further in the past to alleviate this timing constraint. In doing so, though, we run the risk of identifying an outdated set of wealthy families, both because the sampling frame may be outdated and because the income in the new sampling frame may be outdated.

However, we show that we can move the reference year of the frame back with little to no cost in the data quality of the sampling frame. In a comparison of the 2013 SCF list sample frame—which is drawn from a frame of tax year 2011 income (along with 2010 and 2009 income for most families in the 2011 frame)—or a hypothetical list sample frame based on a 2010 income frame (along with 2011, 2009, and 2008 income for most families in the 2010 frame), we are able to show that (1) the impact on the sample composition of using the hypothetical file instead of the actual 2013 SCF sampling file appears to be negligible, on average, and (2) that using the hypothetical file instead of the actual 2013 SCF sampling file does not increase sampling variability.

Change in sample composition?

Because the INSOLE file oversamples records with large incomes (Czajka, et al., 2014; Bryan, 2015) and uses a Keyfitz sampling method, there should be strong overlap between INSOLE files in adjoining years. Of the records in the 2010 INSOLE file—the basis for our hypothetical sampling frame—about 74.5 percent are found in the 2009 INSOLE file and 77.1 percent are found in the 2011 INSOLE file (table 1, panel A). Overall, then, the 2010 INSOLE file should have a similar composition to the 2011 INSOLE file (and to the 2009 INSOLE, as well).

However, using the 2010 file means that we will miss out on the families that are in the 2011 file but not in the 2010 file. To gauge how much this difference in sample composition matters, we compare the mean 2011-2009 income of sampled groups in the 2011 file and in the 2010 file. We begin this exercise by going through the process of selecting a sample with the hypothetical

sampling data—based on wealth rankings derived from the 2010-2008 income panel. Because we have merged 2011 income onto the hypothetical 2010 sampling data, we can also compute the average 2011-2009 incomes—the same years of income that the actual 2013 SCF was sampled on—of the 2010 INSOLE file.

The 2011-2009 incomes of families that would have been sampled by the 2010 SOI file looks nearly identical to the set of families from the 2011 file, by strata (table 2). In each of the seven sampling strata, the difference in mean income is less than five percent. For example, when using the 2011 SOI file and ranking families by their 2011-09 income, the average income of families ranked to stratum six was $6.48 million. The average 2011-09 income of families in the 2010 SOI and ranked by their 2010-08 income was about 4 percent lower at $6.21 million. The impact of moving the sampling frame back to 2010 on the sample composition appears to be negligible, on average, when measuring by total income.[26]

This is true not just for total income but also for key sources of income, especially financial and business income (table 2, panels B and C). Mean 2011-09 financial income of the 2010 stratum sixes ($1.284 million), for example, is about 1 percent higher than the mean 2011-09 financial income of the 2011 stratum sixes ($1.270 million). Mean differences for rent and royalty income, estate and trust income, Schedule C gross income, salary income and, to a lesser extent, capital gain income are also small (appendix table B1). Overall, then, basing the list sampling frame on the INSOLE file from the previous year does not appear to induce a bias in the income composition of the sample.

Change in sampling efficiency?

However, using data from a lagged year to sample may induce more variable rankings, as the sampling data may be less likely to reflect the current economic situation. When families in the 2010 INSOLE file are ranked by their 2010-2008 income and then by their 2011-2009 income, about 85 percent of families in sampling strata two and above remain in the same sampling strata

[26] A mirror exercise with the interquartile ratio (IQR) shows that income variability within strata of the set of families that would have been sampled by the 2010 SOI file and the set of families that was sampled in the 2011 is roughly the same, too. The largest difference is in strata 7 (IQR is about seven percent larger in the new sampling scheme) and strata six (IQR is about four percent smaller); all other strata change three percent or less. A mirror exercise with the standard deviations also yields similar results, though in this exercise income variability in strata 7 is about 25 percent higher in the new sampling scheme.

across years (table 3, diagonals). Though this fraction is large, it also means that about 15 percent of these 2010 families would have been placed in a different strata had the more-recent income data been used to sample.

However, much of this increase in variability can be offset by using a *full* panel to rank families, which will be available for the 2016 SCF and forward. In the past several surveys, the SCF list sampling data contained as many as 22 percent of frame records with missing longitudinal income data; in the 2013 sampling data, there were 20 and 16 percent missing longitudinal income in 2010 and 2009, respectively (table 1, panel B). Part of this missing longitudinal data is due to timing and uncertainty issues, but most were due to other administrative constraints.[27]

The potential benefit of using a full panel is that stratification gets better for about 10 percent of families. To understand how the missing panel data affect classification in the sampling exercise, we randomly set 22 percent of the full 2008-2010 panel data to have missing panel data. The predicted wealth for these records will be based on 2010 INSOLE data only. Generally about 10 percent of families in strata 2 and above are reassigned when using the full panel instead of the dataset with missing panel data (table 4).[28]

Sample weights

Moving back the reference year of the frame by one year will imply a small change to the weighting scheme. The total population in the list sample frame is adjusted to the *survey year* total population by growing the list sample frame population in strata 3 and above by the growth rate of the CPS between the sample frame year and the SCF survey year. In 2016 this meant growing the population by about 2.7 percent instead of 1.5 percent if the usual INSOLE file was used in sampling.

[27] For the past several surveys, there has been a glitch in the matching algorithm between INSOLE data and unedited IRS administrative data, which was discovered during this work. To be clear, basing the sampling frame on a lagged year does not lead to a higher match rate.

[28] Ten percent should also be taken as a lower bound. Because of timing and matching issues in the past several surveys, we have predominantly used IRS administrative tax data for the panel income for the upper strata rather than INSOLE income. The unedited IRS administrative tax data are noisier than INSOLE data, especially for the types of income held by the upper strata families; noisier income data lead to noisier wealth rankings (Kennickell, 2001). We did not try to model this in the table 4 exercise, though.

Re-stratification

In our shift to move the sampling year back we will also be able to use the information from the most recent INSOLE file to update wealth rankings and wealth strata during the field period. Once these data are available, we will be able to use all four years of income data to re-rank and re-stratify the data. Moving families around may be costly for field staff, as effort may need to be shifted "mid-stream", so we propose using this only sparingly. One such proposal would move only those families whose new wealth ranking puts them solidly in a new strata (above the 90th percentile of a higher strata or below the 10th percentile of a lower percentile).[29] Less than two percent of families are re-classified this way in the hypothetical 2010 sampling data.[30]

If re-classification is needed, the weight of each re-classified family will be changed to the median weight of the new strata. This is similar to a different misclassification correction currently in place in the SCF weighting scheme (Kennickell and Woodburn, 1999). We expect that the proposed re-stratification would impact a relatively few cases (as does the current misclassification correction).

The distribution of wealth across years

The SCF sampling process uses a proxy for permanent income—an average of several years of income—to model wealth and sample wealthy families (Kennickell 2001). Predicting wealth from permanent income will help smooth away transitory, idiosyncratic changes in income that may contaminate a wealth prediction based only on annual income.

When *annual* income is used to predict wealth in years 2008, 2009, and 2010, the distribution of wealth at the top strata can be fairly unstable and identifying a strata of the wealthiest families (say, the wealthiest 1 percent or wealthiest 500 families) is difficult. For example, of the families predicted to be in strata seven (the wealthiest sampling strata) in 2008, only 62 percent are predicted to remain in strata seven in 2009 (table 5, top). Less than 75 percent of the next

[29] In fact, in the 2016 sampling frame, no families met this criteria so no families were re-stratified.
[30] Once the fourth year of data is included in the wealth rankings and stratification, about 10 percent of families would change strata. Most of these families, though, experience very small changes in wealth rankings and their updated wealth rank places them near the border between their old strata and their new strata. For example, of the 1,040 families initially ranked into stratum five, 37 are re-classified as stratum four after the fourth year of data are included in the wealth rankings. But only 9 of the 37 are found "solidly" in the stratum 4 distribution (below the 90th percentile of the new stratum 4 distribution).

wealthiest (strata four, five, and six) are predicted to remain in their respective strata in 2009. Note that the size of the population covered by each sampling strata gets smaller as the strata get higher, so movement out of a strata is typically a movement down.

The benefit of using panel data to rank families is evident when comparing family wealth rankings generated from three years of data to the rankings based on just one year of data (table 6). About 20 percent of families ranked in the wealthiest stratum (strata seven) using just the 2010 income are ranked in the next-wealthiest stratum (strata six) when wealth is predicted with three years of income data (2010, 2009, and 2008). These families had unusually high income in 2010—nearly 50 percent higher than the income received in the other two years (not shown). Similarly, about 19 percent of the 2010-only stratum six families are ranked in stratum five when using all three years of data.[31] Again, these families had income in 2010 more than 50 percent higher than usual. In other words, *wealth* rankings generated from one year of data can be contaminated by families with transitorily high *income*.

IV. Increase in Sample Size

By moving the sampling year back we also induce a larger number of sampled families to be SCF-ineligible because of mortality. We propose to boost the list sample modestly and based on an estimated mortality rate in each sampling strata. Mortality depends on many factors, including race, sex, education, and age. Here we use the mortality tables developed by Brown et al. (2002) which take into account these factors. We cannot adjust our sampling data by either sex or race, so we take white males as our base case. We also do not know education level, but we assume that college degree is perfectly correlated with high wealth and adjust strata three through seven for the lower mortality associated with the highly educated.

The average age in each sampling strata is described in the first column of table 7. The second column describes the mortality rate of that average age, the third column describes the high-education adjustment (Brown et al., 2002), the fourth column shows the education-adjusted mortality rate for the average age in each strata, the fifth column shows the number of families sampled in each strata, and the final column describes the proposed new sample size in each

[31] The asymmetry evident in table 4 is due to the fact that strata capture smaller and smaller fractiles as strata increase. There is more of a chance of a strata six family moving to strata five than to strata seven because

strata. Each modification is minimal but, in total, we propose sampling an additional 37 list sample cases. About half of the added cases are in the top three strata.

Response rates in the list sample have generally held constant between 2004 and 2010. In 2013, though, stratum five had a noticeable decline in response, from about 35% in 2004-2010 to about 30% in 2013. Kennickell (1998) describes the determination of strata sample sizes for the SCF. For example, the 2016 sample size depends on the target number of completed cases in each strata (*i*) in for both the 2016 and 2013 SCF ($T_{i,2016}, T_{i,2013}$), the past sample size in each strata ($S_{i,2013}$), and the number of completed cases in the previous SCF ($C_{i,2013}$):

$$S_{i,2016} = \left\{\left[T_{i,2016} * S_{i,2013}/T_{i,2013}\right] + \left[T_{i,2016} * S_{i,2013}/C_{i,2013}\right]\right\}/2$$

The decrease in completed stratum five cases will necessitate an increase in the stratum five sample size.[32] Following Kennickell (1998), we propose to further increase the sample in stratum five by 90 families (table 8).

V. Geographic Overlap between the List Sample and NORC National Sample

The AP sample—drawn by NORC at the University of Chicago—provides a nationally-representative sample of families, and the list sample—drawn from administrative data derived from income tax returns—provides a wealthy oversample. The list sample is selected from the same areas as the national sample for operational expediency. But, there is no guarantee that the geographic distribution of wealthy families overlaps with the geographic distribution of the US families. Earlier work (Frankel and Kennickell, 1995) suggests considerable but not perfect overlap between the distributions; post-stratification by geography in the weighting procedure should help to correct these inconsistencies (Kennickell and Woodburn, 1999). Recent changes in wealth concentration, though, may imply a change in the geographic distribution of wealth.

The national sample is updated each decade by NORC, and the 2010 national sample is the most recent.[33] Sampling areas in the national sample are defined by the hierarchy of cohesive geographic areas: (1) the CSA, (2) the CBSA, or (3) a cluster of small contiguous counties for

[32] The SCF tries to maintain a consistent level of challenge for field interviewers across surveys (Kennickell, 1998). Because stratum five was considerably more challenging in 2013 than in the past, we will increase sample size in order to make the level of challenge closer to that found prior to 2013.

[33] The NORC sample is representative of more than 99 percent of the US population, as some far-lying areas were deemed too expensive to adequately cover.

areas not in a CBSA.[34] In the first stage of the national sampling process, "certainty" areas are selected, which are defined as areas that contain more than 0.5% of the US population.[35] There are 38 certainty areas selected in the 2010 NORC national sample. An additional 88 areas are then selected based on size and other criteria.

Ideally, the list sample would be drawn based on where wealthy families reside, and that set of areas would be identical to the set of areas drawn in the national sample. Thus, in this exercise we draw a geographic sample of wealthy families and compare it to the national sample areas. Replicating the national sample design, we draw certainty areas based on where the wealthiest families cluster (where more than 0.5% of the wealthiest families reside).[36] Similar to Frankel and Kennickell (1995), we focus only on the certainty areas; replicating the second and third stage of the NORC sample design is beyond the scope of this paper.

There are 36 areas selected with certainty when the sampling objective is identifying the geographic distribution of wealthy families. Of the 36, three areas are selected with certainty in the wealth sampling that are not selected with certainty in the national population sampling. Of the three, one area is selected in the national sample—but not with certainty—and one other area has historically been supplemented to the geographic sample. (Please note that due to disclosure concerns, we will not release to areas sampled in the SCF, so we cannot list the geographic areas in this paper.) We add the final remaining area as another supplemental area in the 2016 SCF sampling frame. This final area is also geographically close to other sampled areas, extra travel cost should be minimal.

The geographic clustering of wealthy families is stronger than the geographic population clustering. The top two wealthy areas contain 19 percent of wealthy families but only 13 percent of the population, and the top 10 wealthy areas hold 45 percent of wealthy families but only 33 percent of overall population. Sorted by the population of the national sampling areas, the largest 19 areas hold about 55 percent of wealthy families and about 44 percent of the US population.

[34] Nearly all US families reside in one of the 917 Core-Based Statistical Areas (CBSAs), which are defined by an urban core and the surrounding commuting areas. Larger areas are defined by a Combined Statistical Area (CSA), which is typically a group of several CBSAs; there are 169 CSAs as of 2012.
[35] See http://www.norc.org/Research/Projects/Pages/2010-national-sample-frame.aspx
[36] For this exercise and with comparability to Frankel and Kennickell (1995) in mind, wealthy families are defined as those in the top three sampling strata (strata 5, 6, and 7). Families selected for the top three strata completely cover the top half-percent of the expected wealth distribution and hold about 36 percent of US wealth in the 2013 SCF.

By these measures, geographic clustering of wealthy families is mostly unchanged since Frankel and Kennickell (1995).[37]

VI. Conclusion

For more than 30 years the SCF has sought to provide the best estimates of US household wealth. A key part of the SCF sampling design is the list sample, an oversample of expectedly-wealthy families.

The list sample design has evolved over time. The number of completed list sample cases has increased from about 850 in 1989 to about 1,500 in 2013; the strata cutoffs have shifted from hard dollar values to percentiles of the wealth distribution, and a second (empirical correlation) model was introduced to the sampling process in 1995 SCF; and multiple years of income data were introduced to the sampling process in the 2001 SCF. In each instance of change, the pros and cons were weighed.[38]

We propose a new evolution to the list sample by shifting the base sampling frame income data back one year. The new reference year of the frame data will describe income from three years prior to the survey rather than two year prior.

We believe that this change will come with little to no cost in the data quality of the sampling frame. First, we have showed that the set families that would have been sampled using a sampling frame shifted one year back are observationally equivalent to those that were actually sampled for the 2013 SCF. Second, we show roughly offsetting increases and decreases in sampling variability from this change, relative to the 2013 SCF frame data. Using lagged income data leads to more variability, but using a fuller panel (and one with more panel INSOLE data)

[37] In terms of sampling, differences in geographic clustering of wealthy families from the overall population clustering can lead to under (or over) representing of wealthy families. Because the list sampling process drops areas outside of the areas selected by NORC in the national sample, the weight of the remaining tax records in the list sampling frame are scaled-up by the inverse of selection (at the area level) into the NORC sample. Thus, tax records in areas sampled with certainty in both the wealthy and population samples get the correct second-stage weight, as each should be divided by probability one. Almost two-thirds of the sampled units have the correct weight. For the other one-third, tax records in areas where the probability of selection by wealthy families is higher than the probability of selection by overall population will be oversampled because of the scaling-up of weights. This issue has been present since Frankel and Kennickell (1995); the analysis here suggests that this technique remains effective.

[38] See Kennickell and Woodburn (1992) for 1989 list sample completes, Frankel and Kennickell (1995) for the first description of the empirical correlation model, and Kennickell (2001) for the introduction of multiple years of income data.

relative to the past will lead to decreased sampling variability.[39] Further, the use of the most recent income data—though not in sampling—will help alleviate misclassifications.[40]

The SCF list sample has always been constrained to sample families only from the same areas as the NORC national sample. The survey weights have always adjusted for this inability to sample from the entire INSOLE sampling frame, but increasing wealth concentration (Bricker et al., 2016, Saez and Zucman, 2016) may mean that the adjustment won't work presently. However, we find that the current geographic distribution of wealthy families looks similar to past estimates (Frankel and Kennickell, 1995). We also describe how the 2016 SCF list sample will be augmented by around 100 sampled cases, though the number of completed cases will be the same as in recent years.

[39] The cost of moving back the frame reference year by one year also increases when income follow a random walk. In this case, the most recent income realization would be the best measurement, and would contain all the information needed to predict future income. The four-year panel of income used in this paper is too short to properly identify this time series property, but estimates from a 23 year time series of SOI household income (DeBacker, et al., 2013), and twenty years of PSID data (Baker, 1997; Altonji et al., 2013) indicate that earnings and household income do not follow a random walk.

[40] In the 2016 SCF weighting process, there are actually slightly *fewer* misclassifications than in past years (without re-stratfying), indicating that sampling variability was not increased in 2016.

References

Altonji, Joseph, Anthony Smith, and Ivan Vidangos (2013) "Modeling Earnings Dynamics," *Econometrica*, Vol. 81, No. 4 pp 1395-1454.

Baker, Michael. 1997. "Growth-Rate Heterogeneity and the Covariance Structure of Life-Cycle Earnings", *Journal of Labor Economics*, Vol. 15, No. 2, pp. 338-375.

Bricker, Jesse, Lisa J. Dettling, Alice Henriques, Joanne W. Hsu, Lindsay Jacobs, Kevin B. Moore, Sarah Pack, John Sabelhaus, Jeffrey Thompson, and Richard A. Windle. 2017. "Changes in U.S. Family Finances from 2013 to 2016: Evidence from the Survey of Consumer Finances." *Federal Reserve Bulletin* Vol. 103, no. 3.

Bricker, Jesse, Alice Henriques, Jacob Krimmel, and John Sabelhaus. 2016. "Measuring Income and Wealth at the Top Using Administrative and Survey Data." *Brookings Papers on Economic Activity*, Spring.

Brown, Jeffrey, Jeffrey Liebman, and Joshua Pollet. 2002. "Estimating Life Tables that Reflect Socioeconomic Differences in Mortality," in M. Feldstein and J. Liebman, *The Distributional Effects of Social Security Reform*, University of Chicago Press: Chicago, IL, pp. 447 - 457

Bryan, Justin. 2015. "High-Income Tax Returns for 2012", Statistics of Income Bulletin, Summer, pp. 1-60.

Czajka, John, Amang Sukasih, and Brendan Kirwan. 2014. "An Assessment of the Need for a Redesign of the Statistics of Income Individual Tax Sample" mimeo.

Debacker, Jason, Bradley Heim, Vasia Panousi, Shanthi Ramnath, and Ivan Vidangos. 2013. "Rising Inequality: Transitory or Persistent? New Evidence from a Panel of U.S. Tax Returns." *Brookings Papers on Economic Activity*, Spring: 67–122.

Frankel Martin and Arthur Kennickell (1995) "*Toward the Development of an Optimal Stratification Paradigm for the Survey of Consumer Finances*" mimeo.

Greenwood, Daphne. 1983. "An Estimation of U.S. Family Wealth and Its Distribution from Microdata, 1973." *Review of Income and Wealth* 29, no. 1: 23–44.

Kennickell, Arthur. 1998. "List Sample Design for the 1998 Survey of Consumer Finances", mimeo.

Kennickell, Arthur. 2001 "Modeling Wealth with Multiple Observations of Income: Redesign of the Sample for the 2001 Survey of Consumer Finances", mimeo.

Kennickell, Arthur. 2007. "The Role of Over-sampling of the Wealthy in the Survey of Consumer Finances" mimeo.

Kennickell, Arthur B., and R. Louise Woodburn. 1992. "Estimation of Household Net Worth Using Model-Based and Design-Based Weights: Evidence from the 1989 Survey of Consumer Finances" mimeo.

Kennickell, Arthur B., and R. Louise Woodburn. 1999. "Consistent Weight Design for the 1989, 1992, and 1995 SCFs, and the Distribution of Wealth." *Review of Income and Wealth* 45, no. 2: 193–215.

O'Muircheartaigh, Colm, Stephanie Eckman, and Charlene Weiss. 2002. "Traditional and Enhanced Field Listing for Probability Sampling." In *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section*. Alexandria: American Statistical Association.

Saez, Emmanuel, and Gabriel Zucman. 2016. "Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data." *Quarterly Journal of Economics,* Vol. 131, No. 2, pp. 519-578.

Statistics of Income. 2012. *Individual Income Tax Returns.* Washington, DC: Internal Revenue Service

Tourangeau, Roger, Robert A. Johnson, Jiahe Qian, Hee-Choon Shin, Martin R. Frankel. 1993. "Selection of NORC's 1990 National Sample", mimeo

Table 1. Match rates of INSOLE sampling data in adjoining years

Panel A. 2010 INSOLE match rates to 2009 and 2011 INSOLE files, by strata

Strata	2009	2011
1	0.876	0.887
2	0.754	0.781
3	0.599	0.656
4	0.574	0.608
5	0.624	0.670
6	0.783	0.795
7	0.973	0.917
Total	*0.745*	*0.771*

Panel B. Match rates available in sampling (including match to IRS administrative data)

	Hypothetical sampling data (2010 INSOLE)		2013 SCF sampling data (2011 INSOLE)	
Strata	2009	2011	2009	2010
1	0.928	0.932	0.849	0.897
2	0.992	0.978	0.877	0.898
3	0.995	0.983	0.811	0.836
4	0.994	0.984	0.748	0.780
5	0.997	0.991	0.704	0.751
6	0.997	0.993	0.696	0.756
7	0.995	0.996	0.775	0.800
Total	*0.970*	*0.966*	*0.800*	*0.840*

Note: Panel A describes the match rate between an INSOLE file that describes 2010 income to the files describing 2009 and 2011 income. The strata are generated using 2010 income data, translated into a predicted wealth measure (see section II). Panel B describes the match rate of the 2010 INSOLE to either the adjoining years' INSOLE files (as in panel A) or to unedited IRS administrative data. The second set of results describe the actual match rates used in the sampling of the 2013 SCF. The match rates were lower than necessary in 2013 SCF sampling because of a matching glitch (see section III).

Table 2. Mean 2011-09 income of the 2011 and 2010 INSOLE files, strata

2011 INSOLE file, stratified with 2011-2009 income		2010 INSOLE file, stratified with 2010-2008 income		
Mean total income by strata (thous. $2011)				
Strata	Avg. income	Strata	Avg. income	Ratio
1	37.0	1	39.0	1.05
2	109.5	2	110.7	1.01
3	214.2	3	211.2	0.99
4	422.6	4	410.9	0.97
5	1,234.5	5	1,197.6	0.97
6	6,480.8	6	6,213.6	0.96
7	93,391.7	7	97,199.9	1.04
Mean financial income (thous. $2011)				
Strata	Avg. income	Strata	Avg. income	Ratio
1	0.1	1	0.2	1.20
2	3.4	2	3.4	1.00
3	13.5	3	13.9	1.04
4	32.1	4	33.9	1.06
5	138.6	5	139.0	1.00
6	1,270.1	6	1,283.8	1.01
7	32,084.0	7	34,900.1	1.09
Mean partnership/S-Corp income (thous. $2011)				
Strata	Avg. income	Strata	Avg. income	Ratio
1	0.0	1	0.1	1.97
2	1.6	2	1.9	1.19
3	14.5	3	15.7	1.08
4	69.4	4	68.0	0.98
5	379.1	5	364.9	0.96
6	2,775.6	6	2,600.7	0.94
7	44,671.4	7	43,966.9	0.98

First column describes mean income over 2011-09 period of the 2011 INSOLE sampling file (used in 2013 SCF sampling). The second column describes mean income over 2011-09 period of the 2010 INSOLE sampling file (used in this paper as hypothetical sampling data).

Table 3. Transition across strata when 2010 INSOLE file stratified by 2010-08 income, 2011-09 income

		Stratified with 2009-2011 income panel						
		1	2	3	4	5	6	7
	1	0.98	0.09	0.00	0.00	0.00	0.00	0.00
Stratified	2	0.02	0.89	0.15	0.01	0.00	0.00	0.00
with	3	0.00	0.02	0.79	0.12	0.00	0.00	0.00
2008-2010	4	0.00	0.00	0.06	0.85	0.12	0.00	0.00
income	5	0.00	0.00	0.00	0.03	0.86	0.13	0.00
panel	6	0.00	0.00	0.00	0.00	0.01	0.87	0.13
	7	0.00	0.00	0.00	0.00	0.00	0.00	0.87

Note: 2010 INSOLE file is base file for stratification. INSOLE and unedited IRS administrative data for 2008, 2009, and 2011 are merged with the 2010 INSOLE file to create a nearly complete panel (table 2). In the rare case when the 2010 INSOLE records cannot be matched to past income through either the unedited or INSOLE files, the past income is imputed with current income.

Table 4. Transitions across strata when 2010 INSOLE file stratified by full and incomplete 2010-2008 panel

		Stratified with *full* 2008-2010 income panel						
		1	2	3	4	5	6	7
	1	0.99	0.05	0.00	0.00	0.00	0.00	0.00
Stratified	2	0.01	0.94	0.08	0.01	0.00	0.00	0.00
with	3	0.00	0.01	0.88	0.06	0.00	0.00	0.00
incomplete	4	0.00	0.00	0.04	0.91	0.07	0.00	0.00
2008-2010	5	0.00	0.00	0.00	0.02	0.92	0.08	0.00
income	6	0.00	0.00	0.00	0.00	0.01	0.92	0.10
panel	7	0.00	0.00	0.00	0.00	0.00	0.00	0.90

Note: 2010 INSOLE file is base file for stratification. INSOLE and unedited IRS administrative data for 2008, 2009, and 2011 are merged with the 2010 INSOLE file to create a nearly complete panel (table 2). In the rare case when the 2010 INSOLE records cannot be matched to past income through either the unedited or INSOLE files, the past income is imputed with current income. The *full* 2008-2010 panel refers to these data. The *incomplete* 2008-2010 panel refers to a version of these data with a random 22% of observations set with missing panel data.

Table 5. Transitions across strata when one-year of income used to predict wealth

Panel A: 2009 versus 2008

		Wealth rank based on 2008 income						
		1	2	3	4	5	6	7
Wealth rank based on 2009 income	1	0.96	0.15	0.01	0.00	0.00	0.00	0.00
	2	0.04	0.81	0.24	0.04	0.01	0.00	0.00
	3	0.00	0.04	0.65	0.18	0.02	0.00	0.00
	4	0.00	0.00	0.10	0.73	0.21	0.02	0.00
	5	0.00	0.00	0.00	0.05	0.74	0.25	0.01
	6	0.00	0.00	0.00	0.00	0.02	0.72	0.36
	7	0.00	0.00	0.00	0.00	0.00	0.00	0.62

Panel B: 2010 versus 2009

		Wealth rank based on 2009 income						
		1	2	3	4	5	6	7
Wealth rank based on 2010 income	1	0.96	0.15	0.01	0.01	0.01	0.00	0.00
	2	0.04	0.81	0.23	0.03	0.01	0.01	0.00
	3	0.00	0.04	0.67	0.16	0.02	0.00	0.00
	4	0.00	0.00	0.09	0.75	0.19	0.01	0.00
	5	0.00	0.00	0.00	0.05	0.75	0.22	0.01
	6	0.00	0.00	0.00	0.00	0.02	0.75	0.31
	7	0.00	0.00	0.00	0.00	0.00	0.00	0.69

Panel C: 2010 versus 2011

		Wealth rank based on 2010 income						
		1	2	3	4	5	6	7
Wealth rank based on 2011 income	1	0.96	0.15	0.01	0.00	0.00	0.00	0.00
	2	0.04	0.81	0.24	0.02	0.01	0.00	0.01
	3	0.00	0.04	0.67	0.18	0.01	0.00	0.00
	4	0.00	0.00	0.08	0.75	0.19	0.02	0.01
	5	0.00	0.00	0.00	0.04	0.76	0.21	0.01
	6	0.00	0.00	0.00	0.00	0.02	0.76	0.24
	7	0.00	0.00	0.00	0.00	0.00	0.00	0.73

Note: 2010 wealth prediction based solely on 2010 INSOLE file. Wealth predictions in other years are based on either INSOLE income for that year or unedited IRS administrative data for that year.

Table 6. Transition across strata: 2010-08 income panel versus 2010 income

		Stratified with 2010 income						
		1	2	3	4	5	6	7
Stratified with 2008-2010 income panel	1	0.97	0.12	0.00	0.00	0.00	0.00	0.00
	2	0.03	0.85	0.20	0.00	0.00	0.00	0.00
	3	0.00	0.03	0.74	0.16	0.00	0.00	0.00
	4	0.00	0.00	0.07	0.80	0.17	0.00	0.00
	5	0.00	0.00	0.00	0.03	0.82	0.19	0.00
	6	0.00	0.00	0.00	0.00	0.02	0.81	0.20
	7	0.00	0.00	0.00	0.00	0.00	0.00	0.80

Note: 2010 INSOLE file is base file for stratification. INSOLE and unedited IRS administrative data for 2008, 2009, and 2011 are merged with the 2010 INSOLE file to create a nearly complete panel (table 2). In the rare case when the 2010 INSOLE records cannot be matched to past income through either the unedited or INSOLE files, the past income is imputed with current income.

Table 7. List sample sizes adjusted for mortality

Stratum	(1) Average age	(2) Mortality probability	(3) College ed. adjustment	(4) Adjusted mort. prob.	(5) Sample size	(6) Adjusted sample size
1	41.4	0.25	…	0.25	280	281
2	55.2	0.75	…	0.75	400	403
3	58.3	1	0.62	0.62	580	584
4	58.5	1	0.63	0.63	780	785
5	59.1	1	0.63	0.63	1040	1047
6	60.9	1.2	0.64	0.768	1560	1572
7	63.2	1.5	0.67	1.005	500	505

Average age calculated in 2010 INSOLE data. Mortality probability and college education adjustment to mortality found in Brown et al, 2002. We assume that the college-education mortality adjustment should be applied to the wealthiest set of families, here in strata 3-7 (representing roughly the top 5 percent of the wealth distribution).

Table 8. Change in sample size 2016 list sample, by strata

Strata	(1) T,2016	(2) S,2013	(3) T,2013	(4) C,2013	→	(5) S,2016	(6) Diff(S,2016-S,2013)
1	100	280	100	102		277.3	-2.7
2	165	400	165	167		397.6	-2.4
3	210	580	210	215		573.3	-6.7
4	275	780	275	274		781.4	1.4
5	325	1040	325	278		1127.9	87.9
6	375	1560	375	369		1572.7	12.7
7	50	500	50	53		485.8	-14.2

Note: T_{2016} and T_{2013} describe the target sample size in strata *i* in 2013 and 2016, respectively; S_{2013} represents the past sample size in each strata in 2013 SCF; C_{2013} represents the number of completed cases in the 2013 SCF; and S_{2016} represents the sample size in each strata needed in the 2016 SCF. Calculations based on Figure 9 in Kennickell (1998): *List sample design for the 1998 SCF*.

Appendix A. Details on SCF Sampling Strategy

Data

Since 1992, the Federal Reserve Board (FRB) has contracted the SCF field work to NORC at the University of Chicago and for more than thirty years the SCF has partnered with the Statistics of Income (SOI) Division of the Internal Revenue Service to select a "list" oversample of expectedly wealthy families. The INSOLE data, maintained by SOI, are the main data for the list sample selection. Prior to use, the INSOLE data are statistically edited by SOI to support policy work of Congressional and US Treasury staff (Statistics of Income, 2012).

The INSOLE file from the year prior to the survey (which describes the income from two years prior to the survey) are the main sampling data. Two years of panel data are attached to these records. Often the panel data are from the two previous years of INSOLE data, but sometimes they are from the unedited IRS administrative data. For the 2013 SCF, the sampling data were anchored in 2011, but included 2010 and 2009 panel data on the 2011 INSOLE records.

The INSOLE data used for SCF sampling are anonymized and a great degree of security is involved with this sampling procedure. A formal contract governs the agreement between the FRB (who are responsible for selecting the list sample), SOI, and NORC. None of the three entities will ever know all of the sampling, contacting, and survey information. NORC needs to know the contacting information and collects the survey information but will never know the sampling information. SOI knows the contacting and sampling information but not the survey information. And the FRB knows the sampling and survey information but not the contacting information.

Gross Capitalization Model

The data used to select the 2013 list sample were anchored in 2011 but included 2010 and 2009 panel data on the 2011 records. More weight is given to the income from the most recent tax year (as seen below). These data are read into two models which predict wealth from income. The two models are briefly described in Section II and are described here in detail in unpublished papers on the SCF website, http://www.federalreserve.gov/econresdata/scf/scf_workingpapers.htm.

The exact form of the gross capitalization model in the SCF when selecting the 2013 SCF was:

$$\widehat{wealth}_i^{GC,T} = \frac{max(0,|taxable\ interest_i|)}{ror^{taxable\ interest}} + \frac{max(0,|non\ taxable\ interest_i|)}{ror^{non\ taxable\ interest}} + \frac{max(0,|dividends_i|)}{ror^{dividends}} + \frac{max(0,|rent\ \&\ royalties_i|)}{ror^{rent\&royalties}} + \frac{(|partnerships\ \&\ S{-}corps_i|+|estates\ \&\ trusts_i|)}{(ror^{dividends}+ror^{non\ taxable\ interest})/2} + \frac{(|schedule\ C\ gross\ income_i|+|gross\ farm\ income_i|)}{(ror^{dividends}+ror^{non\ taxable\ interest})/2} + net\ capital\ gains_i + \widehat{house}_\iota\ ,$$

where, there are where there are i=1…N tax units,

$$inc\ concept_i = \frac{1}{2} * inc\ concept_i^{2011} + \frac{3}{10} * inc\ concept_i^{2010} + \frac{2}{10} * inc\ concept_i^{2009},$$

and:

$$ror_i^{inc\ concept} = \frac{1}{2} * ror_i^{inc\ concept,2011} + \frac{3}{10} * ror_i^{inc\ concept,2010} + \frac{2}{10} * ror_i^{inc\ concept,2009},$$

for:

$$inc\ concept_i = taxable\ interest, non\ taxable\ interest, dividends, rent\ \&\ royalties, partnerships\ \&\ S-corps, estates\ \&\ trusts, schedule\ C\ gross\ income, gross\ farm\ income, net\ capital\ gains.$$

The rate of return on taxable interest is based on the Federal Reserve H.15 data series on the AAA corporate bond rate (seasoned issue, all industry). The rate of return on non-taxable interest is based on the H.15 data series on Moody's June rate on AAA state and local 20-year bonds. The rate of return on dividends is based on the S&P dividend price ratio, and the return on rent and royalties is based on the effective yield from a 30-year conventional mortgage from the H.15 data series. The rate of return on businesses, estates, trusts, and farms is estimated to be the mean of the rate of return of taxable interest and dividends. Capital gains are not adjusted for a rate of return.

Predicted home equity is based on finding the median house value within that tax unit's income range from the most recent SCF; the 2010 SCF data were used in selecting the 2013 list sample. Tax units are grouped into those with less than $60,000 in income (in $1989), between $60,000 and $120,000, between $120,000 and $250,000, between $250,000 and $1,000,000, between $1,000,000 and $5,000,000, and greater than $5 million in income.

Table A.1. Predicted home equity for gross-capitalization model

	Median value in 2010 SCF
Less than $60,000 in income ($1989)	$114,140
Between $60,000 and $120,000 in income ($1989)	$354,125
Between $120,000 and $250,000 in income ($1989)	$703,400
Between $250,000 and $1,000,000 in income ($1989)	$1,300,605
Between $1,000,000 and $5,000,000 in income ($1989)	$2,416,087
More than $5,000,000 in income ($1989)	$6,085,780

Empirical Correlation Model

The second model uses the empirical correlation between past SCF wealth and sampling data to predict a wealth ranking in the current sampling data. In selecting the 2013 list sample, the 2010 SCF wealth was linked to the sampling data for the 2010 SCF; these sampling data are the panelized version of the 2008 INSOLE file. A special dispensation granted by SOI allows this link for the purpose of selecting the list sample.

The sampling data contain many sources of income. The first step in the empirical correlation modelling process begins by finding the sampling variables that are most correlated with wealth. The sampling variables can describe income or certain deductions.

The process begins with a simple regression of logged SCF wealth on logged dollar values of sampling data and dummies for positive values of each income type; a stepwise selection process is used to determine which of these variables are most highly correlated with SCF wealth. In a stepwise selection criteria, the most variables most highly correlated with SCF wealth are sequentially added until all highly correlated variables are included; once a variable is added, the process also removes the variables that lose their correlation with wealth once the added variable is included in the model. The criterion for inclusion in the model is a p-value of 0.35. Some theoretically-relevant variables are added even if they are not selected in the stepwise selection process.

Thirty-three income variables in total are selected for the model, along with several geography dummies, marital and filing status, and age variables. These variables are included in a final first step model to find the correlation between SCF wealth and sampling data:

$$\ln(SCF\ wealth_i^{2010}) = \alpha + \beta_L^1 \ln\left(income_i^{1,2008-06}\right) + \beta_D^1 \mathrm{I}\left(income_i^{1,2008-06} > 0\right) + \cdots + \beta_L^{33} \ln\left(income_i^{33,2008-06}\right) + \beta_D^{33} \mathrm{I}(income_i^{33.2008-06} > 0) + X_i^{2008-06}\delta + \varepsilon_i,$$

where $X = [geography, marital, filing, age]$,
and $\ln\left(income_i^{j,2008-06}\right) = \ln(\left|\frac{1}{2} * income_i^{j,2008} + \frac{3}{10} * income_i^{j,2007} + \frac{2}{10} * income_i^{j,2006}\right|)$,
for j=1…33

The $\hat{\alpha}, \widehat{\beta_L}, \widehat{\beta_D}, \hat{\delta}$ vector from this regression model is then applied to the current administrative sampling data (for which the same income variables are available) to get a predicted wealth index, which we denote here as the "empirical correlation" prediction:

$$\widehat{wealth}_i^{ECorr,2013} = \alpha + \hat{\beta}_L^1 \ln\left(income_i^{1,2011-09}\right) + \hat{\beta}_D^1 \mathrm{I}\left(income_i^{1,2011-09} > 0\right) + \cdots + \hat{\beta}_L^{33} \ln\left(income_i^{33,2011-09}\right) + \hat{\beta}_D^{33} \mathrm{I}\left(income_i^{33,2011-09} > 0\right) + X_i^{2011-09}\hat{\delta}.$$

Final rankings

The two predictions are blended together and used to rank the INSOLE families from highest to lowest expected wealth. In the 2013 selection process, the blend was:

$$blend_i^{2013} = \frac{1}{2}\left\{\frac{\widehat{wealth}_i^{ECorr,2013} - median(\widehat{wealth}_i^{ECorr,2013})}{IQR(\widehat{wealth}_i^{ECorr,2013})} + \frac{\widehat{wealth}_i^{GC,2013} - median(\widehat{wealth}_i^{GC,2013})}{IQR(\widehat{wealth}_i^{GC,2013})}\right\}.$$

The IQR() represents the interquartile range. In past years, the $blend_i$ weighted the empirical correlation model more than the gross capitalization model, in part because of the results shown in table 2 and 3 of this paper. The weight was even in the 2013 selection process.

Families in the Forbes 400 and other families who finances are too unique for public data disclosure are removed from the sample.

Sample Selection

A probability proportional to size (PPS) method is used to select the sample. PPS sampling can be described through the following example:

A statistician wishes to select 100 families from a set of 1,000 families. The families are order from 1 to 1,000 and a sampling interval equal to 10 (=1000/100) is computed, which bins off the families into 100 bins of 10 families. Find a random number between 1 and 10; if the number is 6 then select the 6^{th} family, the 16^{th} family, the 26^{th} family, etc… until 100 families are selected.

If each family has a sampling weight associated with it (as the INSOLE data do) then the example changes a bit. Assume that the first seven-hundred and fifty families have a weight of 1 and the next 249 have a weight of 10 and the final family has a weight of 60. Instead of 1,000 total families, the statistician actually picks from a weighted total of 3,400. The statistician still want to select 100 families, so the sampling interval is 34 (=3400/100) and there are 100 bins of 34 families. The families are ordered from highest weight to lowest then the family with weight of 100 is selected with certainty. Draw a random number between 1 and 34, say 31, then select the 31^{st} family (which is the family with weight of 60), then the 62^{nd} family, the 93^{rd} family, etc… until 100 families are selected.

The list sample is selected in a similar fashion, with observations stratified by predicted wealth, and sub-stratified by age and financial income.

Appendix B. Mean income comparison of 2011 and 2010 SOI files

Table B1. Mean 2011-09 income of the 2011 and 2010 SOI files, by strata and income type

2011 SOI data, stratified with 2011-2009 income

2010 SOI data, stratified with 2010-2008 income

Mean salary income (thous. $2011)

Strata	Avg. income	Strata	Avg. income	Ratio
1	30.6	1	32.1	1.05
2	69.9	2	70.1	1.00
3	97.1	3	93.1	0.96
4	130.5	4	124.7	0.96
5	263.6	5	245.0	0.93
6	928.2	6	861.0	0.93
7	3,714.6	7	4,191.9	1.13

Mean capital gain income

Strata	Avg. income	Strata	Avg. income	Ratio
1	0.0	1	0.0	1.08
2	-0.2	2	-0.2	0.93
3	-0.8	3	-0.8	1.02
4	-1.7	4	-1.4	0.79
5	-5.5	5	-4.6	0.83
6	-4.3	6	-5.7	1.34
7	2,794.7	7	1,886.0	0.67

Mean Sch C gross income

Strata	Avg. income	Strata	Avg. income	Ratio
1	0.7	1	0.8	1.21
2	9.3	2	9.5	1.01
3	39.8	3	39.9	1.00
4	94.1	4	94.6	1.01
5	221.8	5	222.8	1.00
6	684.3	6	673.5	0.98
7	8,533.6	7	9,783.9	1.15

Mean estate + trust income

Strata	Avg. income	Strata	Avg. income	Ratio
1	0.0	1	0.0	2.27

2	0.1	2	0.1	1.01
3	0.7	3	0.7	0.95
4	1.7	4	1.8	1.08
5	9.2	5	9.1	0.99
6	163.8	6	141.4	0.86
7	2,549.7	7	2,229.9	0.87

Mean rent + royalty income

Strata	Avg. income	Strata	Avg. income	Ratio
1	102	1	134	1.31
2	3,370	2	3,447	1.02
3	14,093	3	14,237	1.01
4	33,305	4	33,477	1.01
5	92,883	5	95,885	1.03
6	407,986	6	412,575	1.01
7	2,730,877	7	2,774,584	1.02

Note: see table 2 for more details