USING INCOME TO PREDICT WEALTH: AN EXPLORATION WITH PANEL DATA[*]

Jesse Bricker and Alice Henriques

Federal Reserve Board

DRAFT: DO NOT CITE OR CIRCULATE

Abstract

In this paper we explore some of the strengths and limitations of using administrative income data to infer wealth. We show that estimates of top wealth shares inferred using annual income can be very noisy, even though administrative data have excellent coverage at the top. In the 2010 income data, we can estimate a top 1 percent wealth shares anywhere in the range from 43.0 percent to 37.2 percent. We also demonstrate that annual changes in wealth rankings are generally small, even though the composition of top wealth share groups is fairly unstable.

[*] 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 helpful comments from Tom Crossley, Kevin Moore, Steven Pedlow, John Sabelhaus and Jeffrey Thompson. We also thank Arthur Kennickell for inherited knowledge of the SCF sampling process. We have also greatly benefited from sample coordination with the staff at the Statistics of Income, especially Barry Johnson, David Paris, Michael Parisi, Lori Hentz, and Lori Russ.

I. Introduction

The use of administrative income tax data has revived the discussion of income concentration in the past decade. The richness of the data have allowed detailed descriptions of the top of the income distribution and the development new estimates of the income distribution, changes in the income distribution, and income inequality (Piketty, 1999; Piketty and Saez, 2003; DeBacker, Haim, Panoussi, and Vidangos, 2012; Auten, Gee, and Turner, 2013; Guvenen, Kaplan, and Song, 2014). Arguably, this work has allowed the welfare of "the top 1 percent" and the "bottom 99 percent" to become ingrained in US culture and has led to concerns about both equity and macro-economic stability (Piketty, 2014; Stiglitz, 2012).

There are no comparable administrative wealth data, but the share of wealth held by the wealthy one percent can be inferred by capitalizing these same administrative income tax data (Greenwood, 1983).[1] These data have a long time series of cross-sections and are used to describe how wealth concentration has evolved over the past century (Saez and Zucman, 2016). And because tax filing is nearly universal at the top, these data provide excellent coverage at the top of the distribution. However, wealth concentration estimates inferred this way are also quite sensitive to modelling choices (Bricker, Krimmel, Henriques, and Sabelhaus, 2016; Kopczuk, 2015). In this paper we use a four year *panel* of administrative records derived from income tax returns to further describe the qualities of inferred wealth estimates.

First, we show that capitalizing annual income can paint a misleading picture of wealth, as annual income may be contaminated with transitory income changes. These transitory changes may represent a shock unrelated to the balance of the underlying asset or may represent a choice by a family to realize income at an advantageous time. With panel data we construct proxies for permanent and transitory income and estimate wealth concentration including and excluding transitory income changes. Top wealth shares calculated using annual income, contaminated by transitory income, are overstated relative to our measures of [more] permanent income. For example, using only 2010 income data, the top 1 percent are estimated to hold 43.0 percent of wealth in the United States. But using a 2008-2011 panel of income data, the top 1 percent are

[1] The only wealth tax that exists in the U.S. is an estate tax applied at death to very few families. Kopczuk and Saez (2004) use these data to extrapolate the distribution of wealth over a long time-series. Johnson and Raub (2015) provide updated wealth estimates wealth from the estate tax after considering different rates of mortality.

estimated to hold 41.0 percent of wealth. The share of wealth held by the top 0.1 percent in the panel is also considerably lower than the wealth estimates based on one year of income data.

We also describe the *overall* noise associated with wealth concentration estimates inferred from income tax data, and the estimates described here are quite noisy. As shown in our earlier work, wealth inferred from income tax data are also sensitive to assumptions made about rates of return on capital income (Bricker et al., 2016). Together with that work, here we show that the overall sensitivity around inferred wealth concentration estimates can be as much as 15 percent of the estimate itself. In the example above, using market rates of return on capital income and permanent income leads to a 38.7 percent estimate of the top 1 percent wealth share. In other words, wealth inferred from the income tax data in 2010 can support a range of top 1 percent wealth concentration between 37.2 percent and 43.0 percent. Thus, even though administrative income tax data have excellent population coverage of the richest households, top wealth share estimates can be quite noisy.

With this understanding, recent differences in wealth concentration estimates between administrative income data and the Survey of Consumer Finances (SCF) can be put into context. When wealth is predicted from repeated cross sections of administrative income data (Saez and Zucman, 2016), wealth concentration estimates are slightly higher than those from the SCF (Bricker et al., 2016). The SCF oversamples wealthy families from a *panel* of administrative tax records to get precise top-end wealth estimates. By identifying wealthy families through permanent income, the SCF estimates of the top-end wealth shares are based on a more stable set of wealthy families, and as such, wealth measures from the survey are less influenced by transitory income changes.

Membership in the top share groups is sensitive to the use of annual or permanent income. About 20% of the families inferred to be in the top 1 percent with annual income drop out of the top 1 when permanent income is used to infer wealth. Transitory income fluctuations primarily drive these movements in-and-out of top share groups. For example, the typical family that is ranked in the top 1 percent when using one year of income data but ranked outside of the top 1 with permanent income realized 50 percent higher income than usual in that top 1 year. Business income makes up the majority of income for wealthy families and transitory business income and financial income appears to drive much of the flux in estimated wealth rankings at the top.

Noisiness in top group membership, though, is due to how small these top groups are, and wealth rankings for most families are quite stable over our four year panel. For example, about 75% of the top 0.1 percent in 2008 remain in the top 0.1 in 2009, and nearly all of the 25% that fell out remain in the top 1 percent. About 65% of the top 0.1 percent in 2008 are found in the top 0.1 three years later, though only about 55% of the top 0.1 in 2008 are *continually* in the top 0.1 over the next three years.

General stability can characterize families further down the wealth distribution, as the average rank change year-to-year is 1 percentage point for families in the 50th-99th percentiles. Not surprisingly, the estimated wealth distribution from annual data is more stable than the annual income distribution, with half of the annual transitions in-and-out of the top 1 percent described in Auten, Gee, and Turner (2013) or earnings distribution described in Guvenen, Kaplan, and Song (2014).

The rest of the paper is laid out as follows. In section II we describe a set of administrative records derived from income tax data, how to model wealth from these administrative data, and why using measures of permanent income can lead to different wealth estimates than measures of annual income. In section III we describe how well annual income and permanent income data estimate the composition of top wealth groups. In Section IV we describe measures of wealth concentration when using different strategies for estimating wealth from income. Section V concludes.

II. Data and Wealth Models

The data used in this paper are the sampling data for the Survey of Consumer Finances: the Individual and Sole Proprietor (INSOLE) data file maintained by SOI (Statistics of Income, 2012). The SCF is a cross-section survey, conducted every three years by NORC on behalf of the Federal Reserve Board (FRB) and with the cooperation of the Statistics of Income (SOI) at the Internal Revenue Service (IRS). The SCF provides the most comprehensive and highest quality microdata available on U.S. household wealth.[2]

[2] See Bricker, et al (2014) for results from the most recent triennial SCF, and Bricker, et al (2016) for a comparison of SCF wealth estimates to other wealth data sources.

The SCF combines a geographically-stratified and nationally-representative area probability (AP) sample with a list sample (LS), 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.[3] The LS is drawn using a frame of statistical records derived from tax returns, statistically edited for quality by SOI.[4] The LS ensures that the SCF has adequate representation of the upper tail of the wealth distribution and of sparsely held assets.

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 LS selection process depends on inferring wealth from income tax return information from the sampling data file.[5]

The LS sampling frame of statistical records derived from tax returns has historically been augmented to include two past years of returns so that the sampling data are actually a three year panel. For the exercise in the paper, a four year sampling panel based on the 2010 returns file was created, spanning 2011-2008 returns.[6]

About 97 percent of the 2010 returns file can be matched back to a 2009 return, about 96 percent can be matched back to the 2008 tax records, and about 97 can be matched to a 2011 return (not shown). A full panel is found for more than 99 percent of the families in the top 1% of wealthy filers. Most of the unmatched records belong to families in the bottom of the income distribution, and the median 2010 income of the un-matched records is around $9,000; we can infer that the majority of these un-matched records are primarily due to individuals who do not regularly file tax returns.

[3] See Tourangeau, et al. (1993), O'Muircheartaigh et al. (2002), and Pedlow (2012) for more information about the 1990, 2000, and 2010, respectively, NORC national samples.

[4] Prior to use, the data are edited by SOI to support research at the Office of Tax Analysis and the Joint Economic Committee of the Congress (Statistics of Income, 2012). 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.

[5] The models that predict wealth from income are described in more detail in Kennickell (2005).

[6] Moving the SCF sampling year back one year will allow a full sampling panel. The work contained in this paper is used to support moving the sampling year back.

The unit of observation in the sampling 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. The LS sampling process adjusts for this disconnect since the SCF is a survey of families. Without a correction, these multi-filer families would have a disproportionately large chance of being selected for the SCF sample.[7]

Wealth models in the SCF sampling procedure

In order to have adequate representation at the top of the wealth distribution, the SCF sampling process predicts wealth from administrative income tax data. The methods for this estimation and the process of selecting the LS sample has evolved since the current SCF began in 1989, as more refined models for selecting wealthy respondents have been developed, including moving from cross-section to panel-based administrative records in order to better control for transitory income fluctuations (Kennickell, 2005).

The SCF sampling strategy uses two methods of predicting wealth from income. The first is a gross-capitalization model, generated by inflating the tax unit's asset-based income by an asset-specific rate of return and adding a predicted housing value (Greenwood, 1983). The general form of the SCF model is:

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

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).[8]

[7] The sampling file implies that there were 135 million tax units in 2012; Saez and Zucman (2014) estimate there were 160 million filing and non-filing tax units in 2012. The CPS implies about 120 million families in 2012. To account for this in the SCF LS sampling process, the sampling weight of tax units in the frame 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 sampling frame data.

[8] Model details are provided in Appendix A, including rates of return. Income is a weighted average of three years of sampling income. Saez and Zucman (2016) also use a gross capitalization model to predict wealth from SOI income data but rely on a single year of income tax data. In their version, the rate of return for each capital asset type is defined by the ratio of SOI income for type of asset income to the stock of household (and non-profit) assets in the Financial Accounts for each asset type. The end result is that the Saez and Zucman (2014) gross capitalization method allocates wealth according to SOI income and predicted wealth will match the household (and non-profit) wealth in the Financial Accounts. The rates of return used in the SCF are similar to those used by Saez and Zucman (2016) with the exception of the return to interest-bearing assets.

The second model 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.[9] The $\hat{\beta}$vector from this regression model is then applied to the current administrative sampling data to obtain a predicted wealth index:

$$\widehat{wealth_{it}^{ECorr}} = f(\overline{Income}_{it}\,; \widehat{\beta_{T-3}}).$$

Both the empirical correlation and gross capitalization models use the average over multiple years of administrative data in order to identify wealthy individuals, which helps to smooth over the effects of transitory income fluctuations that are especially prevalent for capital incomes and at the top of the distribution. In contrast to the gross-capitalization model, two key differences are that the empirical correlation model allows for a variety of income variables that are not necessarily based on a physical asset and that it allows rates of return to vary across different types of families.

The gross capitalization and empirical correlation models generate two independent sets of wealth indices. Both indices are normalized by subtracting its median and dividing by its interquartile range. The two normalized indices are then blended together and the sampling data are ranked from least wealthy to most wealthy by the blended index.[10]

Seven wealth strata are created by this ordered index. The top one percent are covered by the top four sampling strata while the wealth of filers in the lowest stratum is often comparable to the AP sample. The top 500 families are placed in the top strata (strata seven), and the rest of the

[9] 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.

[10] Typically, the blend is a 50/50 split, although in recent years the split has favored the windex-1 model, due to the strengths discussed in Bricker et al (2015).

units placed into one of the other six wealth strata of increasing expected wealth; the probability of being sampled increases as the value of the strata increases.[11]

Other wealth studies

The SCF uses both the gross capitalization and the empirical correlation models to identify wealthy families from the sampling income data. While the Federal Reserve is the only entity with access to this specific empirical correlation model, the gross capitalization model has been used on several other studies of wealth, and notably proposed in Greenwood (1983) to estimate the wealth concentration in 1973. Greenwood (1983) used a match of the Current Population Survey (CPS) to a dataset of statistical records derived from tax returns. Importantly, the CPS-matched data allowed a look at *family* wealth concentration, rather than tax-unit wealth concentration. In 1973, the top 10 percent of families were estimated to hold nearly 70 percent of net worth, and the top 1 percent held 32.6 percent.

Recent work has applied Greenwood (1983)'s gross capitalization framework to nearly 100 years of SOI data to get long time series estimates of the change in wealth concentration in the United States (Saez and Zucman, 2016).[12] Wealth concentration was very high in the 1920s and early 1930s, before falling in the 1940s, and staying consistently lower through the early 1980s. In this longer time context, the 1973 wealth concentration estimates are shown to be near the historical low. This recent analysis shows wealth concentration has risen again, nearly to the levels recorded in the 1920s. In these estimates, then, wealth inequality has followed a U-shape over the past century.[13] By using only tax data, this work can only give wealth inequality estimates of tax units rather than families.

[11] 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.

[12] The rates of return used in Saez and Zucman (2014) are generated from the ratio of SOI income of asset type *k* to the asset stock in the Financial Accounts of the United States (FA). The SCF gross capitalization model uses market rates of return, which are generally comparable to the Saez and Zucman (2016) rates of return, save the ratio of taxable interest to fixed income assets.

[13] See Bricker et al. (2016) for a critique of the recent increase in wealth inequality measured in Saez and Zucman (2016). Particularly, the entire increase in wealth inequality 2000-2012 is due to increased holdings of fixed income type of assets, and particularly savings and transaction accounts. This portfolio choice is implausible for the top of the wealth distribution. Further, Johnson and Raub (2015) demonstrate that the implied mortality rates in Saez and Zucman (2016) are too low for older individuals, meaning that the wealth of wealthy families in Saez and Zucman (2016) may be overestimated by as much at 10 percent in the year 2007.

Estate-tax data has also be used to estimate wealth concentration since the early 1900s (Kopczuk and Saez, 2004). Here, combining estate tax filings and mortality rates can provide an estimate of wealth concentration even with few families ever filing estate taxes. Importantly, the time series of estimates of the estate tax do *not* show the U-shape in the time series of SOI income data (Saez and Zucman, 2016). Wealth inequality estimated from estate tax data has *not* increased in recent years. Prior to the mid-1980s, though, the roughly 70-year trend in wealth concentration from repeated cross-sections of estate-tax estimates are similar to the repeated cross-sections of wealth predicted with SOI income data in Saez and Zucman (2016).

Estimates both from estate tax and annual income tax data identify expectedly-wealthy families from repeated cross-sections of administrative data and have little to say about dynamics of wealth. However, some household surveys collect wealth information across time. Among household surveys, only the SCF can claim to capture a representative snapshot of the US wealth distribution. The SCF, though, does not typically provide wealth dynamics because it is a rarely conducted as a panel. The recent exception is the 2007-09 SCF panel, which showed significant churning across the wealth distribution (Bricker et al., 2011) and the top 1 percent. About 33 percent of the top 1 percent in the 2007 SCF dropped out of the top 1 percent in the 2009 follow-up (Kennickell, 2011).

Data from the Panel Survey of Income Dynamics (PSID) also shows significant churning in the wealth distribution over time (Pfeffer et al, 2013; Conley and Glauber, 2008). But without an oversample of wealthy families, the PSID estimates generally represent the bottom 95 percent of the wealth distribution – the part of the distribution where most *families* are found, but where less than half of total *wealth* is found.

Why use permanent income instead of transitory income?

Models of economic behavior often use a concept of *permanent* income to explain economic behavior of families (see Meghir and Pistaferri, 2011, for a review). Families make plans with both today and the future in mind and may shift income to either time period in order to smooth consumption (or, alternatively, may access credit markets to do so). This may be especially true for wealthy families, who typically have substantial business income and investment income and can choose when to realize that income.

Aside from permanent income models, families may respond to changes in the tax code and realize income at advantageous tax years (Wolfers, 2015). Thus, one year of income (y) for family i in year t is often some function of a family's permanent income (p) and a transitory component to income (ε):

$$y_{i,t} = p_{i,t} + \varepsilon_{i,t}.$$

The literature often uses a blend of several years of income to proxy for permanent income (DeBacker et al., 2012).

For these reasons, using annual income to model wealth may lead to misleading predictions about the family's stock of wealth. Take, for example, the case of two families with identical $1 million equity holdings, but one family held this amount in 2011 and the other in 2012. Each could expect to receive about $20,000 in dividend income (based on a 2% dividend-price ratio). But the family in 2012 has an incentive to realize more than $20,000 in dividend income because they know that tax rates will rise on dividend income the following year. If that family realized, say, $30,000 in dividends then a capitalization model would predict that family held 50% more equity wealth than the other family, in effect allocating $500,000 in wealth that does not actually exist. Using a blend of three years of dividend income would control for such deviations from reality.

Despite these theoretical reasons for using proxies for permanent income, *a priori*, it is not completely clear that permanent income will yield better wealth predictions than will annual income. After all, if income is a random walk then income today is a sufficient statistic for past income.

Wage income and earnings are often modeled as a random walk in permanent income models so that $y_{i,t} = p_{i,t} + \varepsilon_{i,t}$ and $p_{i,t} = \rho * p_{i,t-1} + \tau_{i,t}$ where $\rho = 1$. Some support for this specification is found in, for example, Meghir and Pistaferri (2004), Topel and Ward (1992), and MaCurdy (1982) but others find that annual earnings cannot proxy for lifetime earnings (Haider and Solon, 2006) and that annual earnings are not a random walk if income profiles are heterogeneous (Guvenen, 2009).

Most of these papers, though, consider only the income that families get by selling labor services, not the passive business or investment income from which top-end families derive their wealth. One paper that explicitly includes all family income is DeBacker et al. (2013), which finds that family income is not a random walk. Thus, for the remained of the paper we take as given that multiple years of income are useful for wealth predictions in the capitalization model.

III. Differences in wealth when predicted by annual or permanent income

Generating estimates of the top 1 percent wealth share first requires a rank ordering of wealth to identify which families are in the top 1 percent. Estimates presented in this section use both our four year panel based on the 2010 INSOLE data (which spans 2008-2011 income years) and the 2010 INSOLE data only. The composition of the top 1 percent, top 0.1 percent, and top 0.01 percent vary across time when annual income is used to rank order families.

Estimating wealth from income depends on *realized income,* which is variable across years. The measures of permanent income in the panel data, though, allows us to smooth away transitory income changes. Year-to-year changes in the annual wealth ranking estimates are often fairly small, but even small rank changes can lead to large changes in the composition of these top groups.

Transitory vs. permanent income

We begin by examining the importance of transitory income changes on wealth estimates. There is ample reason to believe that wealthy families can choose when to realize income, often at advantageous times (Parker and Vissing-Jorgenson, 2012; Wolfers, 2015). Year-to-year income fluctuations also reflect transitory income fluctuations which may be uncorrelated with the value of the underlying asset.

There is no way to account for transitory changes in income when wealth is inferred from one year of income data. But permanent income can be proxied by using the mean of several contiguous years of income, and deviations from the mean are considered transitory income as in Kopczuk, Saez, and Song (2010) and DeBacker et al. (2013). When using only 2010 income data, transitory income pushed about 20 percent of the estimated top 0.01 wealthy families into that wealthiest group (table 1). Similar, though slightly smaller, fractions of families are pushed into the top 0.1 and top 1 percent wealthy groups by transitory 2010 income.[14] Thus, when

[14] Table 1 is also not unique to 2010 income, as wealth predicted from only the 2008, 2009, and 2011 income files produces a nearly identical table.

wealth is estimated from one year of income, as in Saez and Zucman (2016) or Greenwood (1983), we can expect many “false positives” in the top wealth groups. As transitory income is presumably less tied to permanent underlying assets, the expected effect is to inflate on top wealth shares, as the capitalization model will allocate wealth that is not there.

To get a better sense of how one transitorily high income year can distort the wealth ranking estimates, we look at how income varies across time for the families described in table 1. Families ranked in the top 0.01 in all four years, for example, consistently have income around $20 million (table 2). Much of this income comes from businesses and financial assets (not shown).

Families that ranked in the top 0.01 in the 2010 income data but ranked in the top 0.1 in the average of 2008-2011 income realized about $8.9 million in income in 2010, but about $5 million in each of 2011, 2009, and 2008. In other words, the “false positive” top 0.01 families in the 2010 data had income about 80% higher than usual in 2010. Conversely, the set of “false positive” families that ranked in the top 0. 1 in the 2010 income data but ranked in the top 0.01 in the average of 2008-2011 income realized income about 40% lower than usual in 2010 ($5.6 million instead of their usual $9.4 million). Similar patterns are observed for wealth rank changes across other fractiles within the top 1 percent.

On the other hand, families that were ranked consistently in each fractile had very stable income. The average total income of families ranked consistently in the top 0.01, though, only deviates by five, seven, and two percent from the three year average in 2011, 2010, and 2009, respectively.[15] Not surprisingly, families consistently ranked in, say, the top 1, top 0.1 or top 0.01 had lower income volatility than families that moved across these categories. Thus, any increase in income volatility over time can mean that it will become harder and harder to identify permanently wealthy groups of families with just one year of administrative income data.

Much of the volatility in rankings is due to volatility in financial and business income (table 3). Here, our measure of income volatility is the coefficient of variation (CV) of income (and types of income) within families across the four year panel.[16] The probability of a family in the

[15] A similar pattern is observed for financial income.

[16] The CV is the ratio of the standard deviation of a family’s income across the panel to their mean family income across the four years.

top 1 percent being mis-classified when using the only the 2010 data increases seven percent with a one standard deviation increase in taxable interest income or in dividend income volatility, and increases 9 percent with a one standard deviation increase in business income. The other main sources of family income are much less correlated with a change in wealth ranking classification. (PUT SD INTERPS IN TABLE IN []).

Stability inside and outside the top?

We show above that top predicted wealth share group membership can be quite unstable, with about 20% of *annual* top fractile group members changing group status when *permanent* income is used to predict wealth. But the families that fall out of the top 0.1 usually fall to the top 1 percent (excluding the top 0.1 percent). For families in the top 1 percent, 85 percent of the ranking changes across years are 0.15 percentage points or less (not shown). And of the nearly 20 percent of families that fall out of the top 1 percent, most fall just to the 98th percentile.

Outside of the top groups, we find that these predicted wealth rankings are quite stable across time. Among families ranking between the 50th and 99th percentiles, the average annual change in wealth rank is one percentage point and the interquartile range is about 2.5 percentage points (not shown).

And despite the instability of annual grouping, the persistence of top group membership is greater than that seen in income data. From one year to the next, between 80 and 82 percent of the top 1 percent families stay in the top 1 percent (table 4, left-most columns of panel a). Two years later, 76 to 77 percent of top 1 families in 2008 and 2009, respectively, are still in the top 1 (though not necessarily continuously), and 73 percent of families in the top 1 in 2008 are found in the top 1 percent three years later. In general, this non-contiguous persistence estimate is consistent with a similar estimate from the 2007-09 SCF panel where about 67 percent of the top 1 percent in the 2007 SCF remained in the top 1 percent in 2009 (Kennickell, 2011).

If we define persistence as remaining in a group consistently, then the numbers decrease. Only about 64 percent of families in the top 1 in 2008 are found in the top 1 percent three years later (table 4, right-most columns of panel a).

There is less persistence within the top 0.1 percent of wealth holders. From one year to the next, between 73 and 77 percent of the top 0.1 families remained in the top 0.1, and about 65 percent of the 2008 top 0.1 percent are found in the top 0.1 in 2011, though only about 55 percent of the 2008 top 0.1 percent are *continuously* found in the top 0.1 for those three years (table 4, panel b).

The persistence at the top of the wealth distribution is much stronger than that of the income distribution, where about 60 percent of the top 1 percent in the income distribution remain in a subsequent year, and about 48 percent remain for two subsequent years (Auten et al., 2013).[17]

IV. An application: wealth concentration estimates

Using a panel of administrative income data to classify families into top share groups allows permanent measures of income to be used in the wealth prediction. Estimated wealth concentration is significantly lower when a set of permanently wealthy families are identified. The share of wealth held by the top 1 percent in 2010 is 43.0 percent when only 2010 data are used to predict wealth, but is 41.0 percent when a blend of the 2011-2008 data are used (table 5, panel a). Similarly, the top 0.1 percent hold 18.1 percent of wealth when four years of data are used to rank but are predicted to hold 19.6 percent when only 2010 income data are used (table 6, panel b).

But model parameters also help define these wealth share estimates, a point that we allude to in earlier work (Bricker et al., 2016). The rates of return on capital income thus far have followed the work of Saez and Zucman (2016), where returns are estimated by the flow of income recorded in the INSOLE data relative to the stock of assets recorded in the Financial Accounts of the United States (FA). The second set of columns in table 5 describe estimated wealth shares when rates of return on capital income are based on market rates, such as the 10-year Treasury yields, corporate bond yields, average mortgage interest rates, Moody's rates on state and local Aaa bonds, and dividend-price ratios. Wealth concentration estimated with these returns is muted relative to those used in the first column of table 5. Using the market rates of return, the top 1% share is 38.8 percent using 2010 income alone and 37.2 percent using 2008-2011 income, compared to 43.4 and 41 percent, respectively.

[17] Auten et al., (2013) use 2000-2005 income data; we can confirm these total income estimates in the 2009-2011 data.

V. Discussion

Discussions of income and wealth concentration often focus on the share of wealth held by the top 1 percent. Administrative income data allow estimates of cross-sectional top income shares, though they typically estimate the share held by the top tax units rather than the top families. There is no comparable administrative wealth data, so estimating the US wealth distribution has typically been undertaken by surveys (Bricker et al., 2014; Pfeffer et al, 2014). Predicting wealth from administrative income data holds the promise of being able to estimate the share of wealth held by the top 1 percent, or even the top 0.1 percent and top 0.01 percent (Saez and Zucman, 2016), in a different way.

Wealth concentration estimates from these inferred wealth data are fairly noisy despite the use of administrative data with excellent coverage of the top end. Shown here, top 1 percent wealth share estimates in 2010 range from 37.2 percent to 43 percent depending on whether permanent income is used in place of annual income and whether market rates of return on capital are used in place of alternate rates of return. The share of wealth held by the top 0.1 percent of wealth holders is also noisy, ranging from 16.4 to 19.6 percent in 2010, depending on model inputs.

Past work shows that estimation of top wealth shares can be held back by restrictive models (Bricker et al., 2016; Kopczuk, 2015). Inferring wealth from a single cross section of income can be a tenuous exercise because transitory income flows can lead to misrepresentations of the permanent wealth stock. Here we show how panel administrative income data can be used to predict wealth from permanent income instead. In the panel data, the wealthy top 0.1 percent is a fairly volatile group. Only 55 percent of the wealthy top 0.1 percent persist in the top 0. 1 for all four years of the panel data.

Wealth concentration estimates utilizing cross-sections of administrative income data show both a higher level and higher growth in recent years relative to survey data (Saez and Zucman, 2016). Much of this difference is due to measurement differences between administrative and survey data (Bricker et al., 2016). That said, some of it may be due to how wealthy families are identified: whether in repeated cross sections of administrative income data and using the gross capitalization model (Saez and Zucman, 2016) or from a panel of administrative income data used in conjunction with the SCF, relying on a blend of the gross capitalization and empirical correlation model (Bricker et al., 2016). This section shows that using repeated annual cross sections to classify families into top share groups leads to higher wealth share estimates than using a panel. This divergence can help frame the remaining gap between the adjusted SCF and administrative data estimates discussed in Bricker et al (2016) that were more pronounced for the top 0.1%.

The short panel of data here does not allow us to comment on whether an increase in income volatility at the top (see, for example, DeBacker et al, 2011) will lead to these biases increasing over time.

These findings are important for the SCF sampling process (Bricker and Henriques, 2016). The sampling data for the oversample are the Individual and Sole Proprietor (INSOLE) file maintained by the Statistics of Income (SOI) division of the Internal Revenue Service. The data used here are a three year panel of the 2011 INSOLE data, which we have accessed to test the sampling benefits of using a full panel, rather than an incomplete panel that has historically been used.[18] In these data, more than 99.7 percent of the top 1 percent can be linked across a three year panel.

Though we describe noisy top share wealth estimates and noisy top group membership, the panel data also show that wealth rankings for most families are quite stable over our four year panel. For example, about 75% of the top 0.1 percent in 2008 remain in the top 0.1 in 2009, and nearly all of the 25% that fell out remain in the top 1 percent. About 65% of the top 0.1 percent in 2008 are found in the top 0.1 three years later, though only about 55% of the top 0.1 in 2008 are *continually* in the top 0.1 over the next three years.

This general stability can also characterize families further down the wealth distribution, as the average change year-to-year is 1 percentage point for families in the 50th-99th percentiles. Not

[18] An emerging challenge of running the SCF is the increasing difficulty in completing the survey by the end of the survey field period. Over time, families are harder to locate and more reluctant to divulge sensitive information. Typically, the SCF oversample has been based on INSOLE data available just as the SCF is going into the field, leading to delays in completing the survey on-time. The wealthy families are especially hard to contact, meaning that first contact often does not happen until many months into the field period. The 2016 SCF, though, will be sampled from the INSOLE data from the year prior. Thus, the correlation of wealth over time at the top is especially important to understand for SCF sampling.

surprisingly, the estimated wealth distribution from annual data is more stable than the annual income distribution, with half of the annual transitions in-and-out of the top 1 percent described in Auten, Gee, and Turner (2013) or earnings distribution described in Guvenen, Kaplan, and Song (2014).

References

Auten, Gerald, Geoffrey Gee, and Nicholas Turner 2013 "Income Inequality, Mobility, and Turnover at the Top in the US, 1987-2010." *American Economic Review*, 103(3): 168-72.

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* (forthcoming).

Bricker, Jesse, and Alice Henriques. 2016. "Updates to the SCF Sampling Process," mimeo.

Bricker, Jesse, Lisa J. Dettling, Alice Henriques, Joanne W. Hsu, Kevin B. Moore, John Sabelhaus, Jeffrey Thompson, and Richard A .Windle. 2014**. "**Changes in U.S. Family Finances from 2010 to 2013: Evidence from the Survey of Consumer Finances**,"** *Federal Reserve Bulletin*, 100(4): 1-40. (September)

Bricker, Jesse, Brian Bucks, Arthur B. Kennickell, Traci L. Mach, and Kevin B. Moore (2011). "Surveying the Aftermath of the Storm: Changes in Family Finances from 2007 to 2009," Finance and Economics Discussion Series 2011-17. Board of Governors of the Federal Reserve System (U.S.)

Conley, Dalton, and Rebecca Glauber, 2008. "Wealth Mobility and Volatility in Black and White," Center for American Progress Working Paper.

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*, 67-122. (Spring)

Greenwood, Daphne. 1983. "An Estimation of US Family Wealth and its Distribution from Microdata" *Review of Income and Wealth* (March) pp 23-44.

Guvenen, Fatih, Greg Kaplan, and Jae Song "The Glass Ceiling and the Paper Floor: Gender Differences Among Top Earners, 1981-2012," mimeo.

Johnson, Barry and Brian Raub. "How Much Longevity can Money Buy? Estimating Mortality Rates for Wealthy Individuals," mimeo.

Kennickell, Arthur 2011. "Ponds and Streams" *Finance and Economics Discussion Series*, 2009-13.

Kennickell, Arthur 2005. "The Good Shepherd: Sample Design and Control for Wealth Measurement in the Survey of Consumer Finances," *mimeo*.

Kopczuk, Wojciech. 2015. "What Do We Know About the Evolution of Top Wealth Shares in the United States?" *Journal of Economic Perspectives*, 29(1): 47-66. (Winter)

Kopczuk, Wojciech, Emmanuel Saez, and Jae Song. 2010. “Earnings Inequality and Mobility in the United States: Evidence from Social Security Data since 1937,” *Quarterly Journal of Economics*, Vol. 125, No. 1, pp. 91-128.

O'Muircheartaigh, Colm, Stephanie Eckman, and Charlene Weiss. 2002. “Traditional and Enhanced Field Listing for Probability Sampling,” *American Association for Public Research 2002: Strengthening Our Community - Social Statistics Section*, 2563-2567.

Parker, Jonathan A., and Annette Vissing-Jorgensen. 2010. "The Increase in Income Cyclicality of High-Income Households and Its Relation to the Rise in Top Income Shares." *Brookings Papers on Economic Activity*, 1-55. (Fall)

Pfeffer, Fabian, Seldon Danzinger, and Robert F. Schoeni, 2014. “Wealth Levels, Wealth Inequality, and the Great Recession,” Russell Sage Foundation Working Paper.

Piketty, Thomas. 2014. *Capital in the 21st Century*. Harvard University Press. (April)

Piketty, Thomas, and Emmanuel Saez. 2003. “Income Inequality in the United States, 1913-1998,” *Quarterly Journal of Economics*, 118(1): 1-39.

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

Stiglitz, Joseph E. 2012. *The Price of Inequality: How Today’s Divided Society Endangers Our Future*, W.W. Norton. (June)

Saez, Emmanuel, and Gabriel Zucman. 2016. “Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data,” Forthcoming, *Quarterly Journal of Economics*. Accessible version: National Bureau of Economic Research Working Paper 20625. (October 2014)

Wolfers, Justin. 2015. “The Gains from the Economic Recovery are Still Limited to the Top One Percent” http://www.nytimes.com/2015/01/28/upshot/gains-from-economic-recovery-still-limited-to-top-one-percent.html.

Table 1. Change in wealth percentile strata: 2010 versus 2011-08 rankings

		Ranking with 2010 income				
		Bottom 95	90-95	99-99.9	Top 0.1	Top 0.01
Ranking with 2011-08 income	Bottom 95	0.99	0.16	0.00	0.00	0.00
	90-95	0.01	0.81	0.15	0.00	0.00
	99-99.9	0.00	0.03	0.84	0.18	0.00
	Top 0.1	0.00	0.00	0.02	0.80	0.20
	Top 0.01	0.00	0.00	0.00	0.02	0.80

Note: Column ranking is based on only 2010 returns data only, row ranking is based on 2011-08 panel data described in Section II. Columns sum to 1 (though may not due to rounding). Table describes where families with wealth imputed from only 2010 income data would be ranked when wealth imputed using average of 2011-08 income data. For example, in the last row, of families ranked in top 0.01 of wealth from 2010 income, 80 percent are ranked in top 0.01 when the average of their 2011-08 income data are used to impute wealth.

Table 2. Mean income in fractiles of top 1 percent, by ranking change

			Wealth ranking based on 2010 data			
		Avg. income...	95-99	99-99.9	99.9-99.99	Top 0.01
Wealth ranking based 2011-08 data	95-99	2011-08	500,000	800,000		
		2011	500,000	700,000		
		2010	800,000	1,500,000		
		2009	400,000	600,000		
		2008	500,000	600,000		
	99-99.9	2011-08	800,000	1,300,000	1,800,000	1,940,000
		2011	700,000	1,300,000	1,600,000	<100,000
		2010	600,000	1,600,000	2,800,000	7,450,000
		2009	700,000	1,200,000	1,500,000	<100,000
		2008	1,000,000	1,300,000	1,500,000	200,000
	99.9-99.99	2011-08	2,100,000	2,500,000	3,900,000	6,000,000
		2011	2,100,000	2,400,000	3,800,000	5,400,000
		2010	600,000	1,900,000	4,100,000	8,900,000
		2009	1,900,000	2,450,000	3,600,000	5,100,000
		2008	3,900,000	3,400,000	4,000,000	4,600,000
	Top 0.01	2011-08	10,200,000	10,000,000	9,400,000	20,000,000
		2011	5,800,000	7,800,000	8,700,000	18,000,000
		2010	400,000	1,600,000	5,600,000	21,000,000
		2009	12,700,000	8,000,000	9,500,000	19,000,000
		2008	21,700,000	22,400,000	13,700,000	21,000,000

Note: In dollars. Means rounded to nearest $100,000. Columns show families ranked by wealth predicted from 2011-08 panel data of returns, rows show families ranked by wealth predicted from 2010 returns data only. For example, families ranked in top 0.01 percent in 2011-08 panel data and ranked in top 0.01 in the 2010 data have mean income in the four year panel of about $20 million, and have income in 2010 data of about $21 million. Families ranked in top 0.01 in 2010 but ranked in 99.9-99.99th percentiles in 2011-08 panel data have mean 2011-08 income of about $6.0 million but income in 2010 of about $8.9 million.

Table 3. Income volatility and change in wealth ranking

Variable	Coeff.
cv(salaries)	-0.031
	(0.004)
cv(taxable interest)	-0.117
	(0.005)
cv(non-taxable interest)	0.0005
	(0.003)
cv(dividends)	-0.054
	(0.004)
cv(rent+royalties)	0.032
	(0.004)
cv(partnerships + S-corps)	-0.117
	(0.004)
cv(Sch. C net income)	-0.0001
	(0.0001)
Obs.	47,914
R^2	0.05

Note: table shows estimates from OLS regression of stable wealth rankings on income volatility. Families have stable wealth rank group when they have the same wealth rank group in the 2010-only data and 2010-08 panel data (e.g., table 2). Income volatility is measured by the coefficient of variation.

Table 4. Persistence of rankings across years

Panel A: Top 1 percent

	Non-continuous			Continuous		
	2009	2010	2011	2009	2010	2011
2008	0.80	0.76	0.73	0.80	0.71	0.64
2009	…	0.82	0.77	…	0.82	0.73
2010	…	…	0.82	…	…	0.82

Panel B: Top 0.1 percent

	Non-continuous			Continuous		
	2009	2010	2011	2009	2010	2011
2008	0.73	0.68	0.65	0.73	0.62	0.55
2009	…	0.77	0.71	…	0.77	0.65
2010	…	…	0.77	…	…	0.77

Sources: 2011-2008 panel of returns. Wealth rankings are described in Section II and are based on annual income data, not panel income data. The columns describe the fraction of families that remain in the top fractile in subsequent years. The first set of columns describe how many families remain in the top fractile in a given year, whether or not they persist in all years. The second set describes the fraction of families that remain in the top fractile continually.

Table 5. Wealth concentration estimates, by annual vs. permanent income and by model parameters

Panel A. Top 1 percent

Model 1		Model 2	
2011	43.4	2011	38.8
2010	43.0	2010	38.7
2009	41.9	2009	37.1
2008	40.2	2008	37.8
2011-08	41.0	2011-08	37.2

Panel B. Top 0.1 percent

Model 1		Model 2	
2011	19.0	2011	17.0
2010	19.6	2010	17.8
2009	19.0	2009	16.7
2008	17.9	2008	17.0
2011-08	18.1	2011-08	16.4

Note: wealth predicted from 2011, 2010, 2009, and 2008 returns data, and 2011-08 panel data of returns.
Model 1 uses a capitalization model with rates of return as in Saez and Zucman (2016).
Model 2 uses a capitalization model with market rates of return as in Bricker and Henriques (2016).
Models for predicting wealth from income described in Section II and Appendix A.

Appendix A. Details on SCF Sampling Strategy

The Survey of Consumer Finances (SCF) begins with a traditional household-level Area Probability (AP) sampling frame, and then supplements that sample using administrative data derived from tax records.[19] The administrative data is used primarily to develop the so-called "List" sample of wealthy families, which makes it possible to overcome the problems with thin samples and unit non-response among wealthy families that plague traditional random samples.[20] The List sample is drawn from tax records based on *predicted* wealth, where those predictions are based on income and other observable family characteristics. This section discusses in detail the methodology of the SCF List sample approach, including the strengths and weakness of various models of wealth prediction, and how these models can affect conclusions about measured wealth concentration.

Sampling Overview

The SCF List sampling strategy uses two methods of predicting wealth from income. The first is a gross-capitalization model, generated by inflating the tax unit's asset-based income by an asset-specific rate of return and adding a predicted housing value (Greenwood, 1983). The general form of the SCF model is:

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

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 $\epsilon(0,1)$ and K=6.

The second model 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:

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

[19] The appendix to Bricker et al (2014) provides a high-level overview of the SCF sampling strategy, along with some summary statistics on participation in the Area Probability and List samples. This material is also covered in depth in the FEDS Working Paper version of this paper, Bricker et al (2015).

[20] See, for example, Sabelhaus et al (2015).

The matrix of sampling income, ($\overline{Income}_i^{T-5}$), includes many more types of income than the gross capitalization model, not necessarily based on a physical asset, allows rates of return to vary across different types of families, and permits the inclusion of some basic demographic data.[21]

The $\hat{\beta}$vector from this regression model is then applied to the current administrative sampling data to obtain a predicted wealth index:

$$\widehat{wealth}_i^{ECorr} = f(\overline{Income}_i\ ;\hat{\beta}).$$

Both the empirical correlation and gross capitalization models use multiple years of administrative data in order to identify wealthy individuals, which helps to smooth over the effects of transitory income fluctuations and distinguish the permanently wealthy from those who happen to realize a very high, but transitory, income in a given year. These shocks are especially prevalent for capital incomes and at the top of the distribution.

Sampling Data

The SCF combines a standard nationally-representative area probability (AP) sample with a “List” sample derived from administrative data based on information from tax returns.[22] The List sample is drawn using statistical records derived from tax returns at the Statistics of Income (SOI) Division of the Internal Revenue Service.[23]

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 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-section to panel-based administrative records in order to better control for transitory income fluctuations. The INSOLE data,

[21] 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.

[22] See O’Muircheartaigh et al. (2002) for more information about the NORC national sample.

[23] At the time the sample is drawn, the most recent complete administrative data are those from two years prior to the survey year. The sample includes individual and sole proprietorship tax filings from the IRS administrative tax (see Statistics of Income, 2012).

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 IRS administrative tax 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.

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 6th family, the 16th family, the 26th 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 31st family (which is the family with weight of 60), then the 62nd family, the 93rd family, etc… until 100 families are selected.

The List sample is also selected by a probability proportional to size (PPS) sampling method, stratifying by the seven wealth strata, sub-stratified by age and financial income, with the probability of selection increasing in each stratum.[24] In total, about 5,100 List sample cases are selected; the majority are from strata that capture the top one percent of expected wealth.

Wealthy families are much less likely to respond to a survey (Sabelhaus et al., 2015) and response rates in the List sample vary across strata in an expected manner. The response rate in the wealthiest SCF stratum is around 12 percent, increasing to about 25 percent in the second-wealthiest stratum, 30 percent in the third-wealthiest stratum, 40 percent in the fourth- and fifth-wealthiest and then about 50 percent in the two least-wealthy strata. These response rates are considerably lower than the roughly 70 percent response rate observed in the SCF AP sample.

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

$$\widehat{wealth}_i^{GC,T} = \frac{max(0,|taxable\ interest_i|)}{ror^{taxable\ interest}} + \frac{max(0,|non\ taxable\ interest_i|)}{ror^{non\ taxable\ interest}} + \frac{max(0,|dividends_i|)}{ror^{dividends}} +$$

$$\frac{max(0,|rent\ \&\ royalties_i|)}{ror^{rent\&royalties}} + \frac{(|partnerships\ \&\ S{-}corps_i|+|estates\ \&\ trusts_i|)}{(ror^{dividends}+ror^{non\ taxable\ interest})/2} +$$

$$\frac{(|schedule\ C\ gross\ income_i|+|gross\ farm\ income_i|)}{(ror^{dividends}+ror^{non\ taxable\ interest})/2} + net\ capital\ gains_i + \widehat{house}_\imath\ ,$$

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

[24] Within the seven strata there are nine financial income sub-strata and four age sub-strata. 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.

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

and:

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

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.

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 (Table A.1). 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. The weight was even in the 2013 selection process. With the blended ranking, the sampling data are ordered from least wealthy to most wealthy. Seven wealth strata are created; the wealth of filers in the lowest stratum is often comparable to the AP sample while the top four strata fully cover the top one percent.

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

Model Comparisons

Both the gross capitalization model and the empirical correlation model predict wealth from administrative income. Using the administrative SCF sampling data, it can be shown that the empirical correlation and gross capitalization approaches often disagree on predicted wealth rankings at the very top. The empirical correlation approach and using multiple years of data generate lower predicted top capital income shares, and thus by construction, lower predicted top wealth shares, than gross capitalization alone. Indeed, the differences in predicted top 0.1 percent shares for both wealth and capital income are larger than the residual gaps for the top 0.1 percent identified in the main body of the paper.

Using administrative income records to identify high *wealth* families requires strong assumptions both about the link between taxable income and wealth and about the distribution of wealth components that have no taxable income. To identify the wealthy from income tax data alone, one must rely on annual capital income to infer wealth through the gross-capitalization approach (Greenwood, 1983; Saez and Zucman, 2016).

However, only about half of assets can be inferred from a tax return.[25] When using tax return data, then, the value of these assets must be estimated and benchmarked to aggregate data. The most important "middle-class" assets, like housing and pensions, are typically not included on an income tax return. Saez and Zucman (2016) combine what information can be gathered from tax returns, for example, property tax deductions and current pension payments, with external data, like the SCF, to estimate these types of asset holdings for each tax unit.

Further, the annual *capital* income that is used to estimate wealth from the tax return also has permanent and transitory components. The variance and cyclicality of transitory income has also increased at the top end in recent years (Parker and Vissing-Jorgenson, 2010; Guvenen, Kaplan, and Song, 2014); capital income typically makes up a larger portion of these families' income.[26] An example of this increased variance is seen in the choice of many high-end families to receive capital income in the 2012 tax year in response to increased rates in 2013; predicting wealth using this one-year snapshot will overstate top wealth shares (Wolfers, 2015).

The wealth rankings of each model, using the SCF approach to gross capitalization, are compared here. About 89 percent of families that are predicted to be in the bottom 90 percent in the gross capitalization model are also predicted to be in the bottom 90 percent in the empirical

[25] See Saez and Zucman (2016) Appendix Table A3.
[26] Castaneda, Diaz-Gimenez, and Rios-Rull (2003) look at the dynamic relationship between income and wealth from a theoretical perspective, in the context of a lifecycle model calibration exercise.

correlation model (Table A.2). Looking at finer rankings within top 10 percent, though, there are considerable differences.

Table A.2. Impact of Ranking Top End Families by an Alternate Model

		Correlation Model Percentile				
				(Top 1)	(Top 0.1)	(Top 0.01)
		Bottom 90	90-99	99-99.9	99.9-99.99	99.99+
	Bottom 90	0.89	0.10	0.01	0.00	0.00
Gross-capitalization Percentile	90-99	0.20	0.48	0.28	0.04	0.00
	(Top 1) 99-99.9	0.05	0.22	0.48	0.23	0.02
	(Top 0.1) 99.9-99.99	0.03	0.10	0.31	0.46	0.10
	(Top 0.01) 99.99+	0.01	0.03	0.11	0.39	0.47

Notes: Rows sum to 1. Table describes where a family ranked in gross capitalization model would be ranked in the empirical correlation model. For example, in the last row, of families ranked in top 0.01 percentile in the gross capitalizations model, 1 percent of families are ranked in the bottom 90 percentiles by the correlation model, 3 percent are ranked between the 90-99th percentiles by the correlation model, 11 percent are ranked between the 99th-99.9th percentile by the correlation model, 39 percent are ranked between the 99.9th and 99.99th percentile by the correlation model, and 47 percent are ranked in the top 0.01 percent by the correlation model. Source: 2011 INSOLE data, supplemented with two years of INSOLE or IRS administrative tax panel data.

Within the top 10 percent, slightly less than half of records ranked by the gross-capitalizations model are ranked in the same percentile in the correlation model. Only 47 percent of those ranked in the top 0.01 percent in the gross-capitalizations model are also ranked in the top 0.01 percent in the correlation model. The agreement is at a similar level for the top 0.1 percent (but not in the top 0.01), the top 1 percent (but not the top 0.1), and the top 10 percent (but not the top 1 percent): only 46, 48 and 48 percent, respectively, of those ranked by the gross-capitalizations model are ranked similarly by the correlation model. And viewed another way, 41 percent of families ranked in the top 0.1 percent by the gross capitalizations model are not ranked in the top 0.1 by the correlation model.[27]

Often, the disagreement between the two models in terms of ranking is not large. Of the 53 percent of records ranked in the top 0.01 percent by the gross-capitalizations model that are *not* similarly ranked by the correlation model, 39 percentage points are in the top 0.1 percent (excluding the top 0.01 percent) and only 4 percentage points are out of the top 1 percent when ranked by the correlation model. These classification disagreements are at very fine levels. But often the case for using administrative data are that the sample size allows for the identification of these "top 0.01 percent" or "top 0.1 percent" families (Saez and Zucman, 2016). The results in Table A.2 indicate that such identification is not clear.

[27] Further, the amount of wealth held by families in disagreement between the models is substantial. About 54 percent of the wealth of families ranked in the top 0.1 percent by the gross capitalizations model is held by families ranked below the top 0.1 percent by the regression model.