Abstract:

In this study we aim to provide new estimates of how common it is for a person to achieve income/earnings at the very top of the distribution during their adult lifetime. We expand upon a long-standing literature that has focused on the bottom of the income distribution, i.e. on poverty dynamics, and provide improved estimates of the dynamics of affluence by using survey and administrative data including the Panel Study of Income Dynamics, the Survey of Income and Program Participation linked to Social Security Administration earnings records, and the IRS 1987 Family Panel. Our findings suggest that membership in the upper echelons of the income/earnings distribution has become less accessible over time, suggesting that the barrier between elites and non-elites has become less permeable and that class structure at the top end has become more rigid.

*Lloyd Grieger is an Assistant Professor of Sociology and Global Affairs and a faculty affiliate of the Center for Research on Inequalities and the Life Course at Yale University. Fabian Pfeffer is a Research Assistant Professor at the Survey Research Center and Co-Investigator of the Panel Study of Income Dynamics at the University of Michigan.

Introduction:

“By the time we’re 60, three in five Americans will have spent at least a year at the bottom. If that sounds like a surprisingly high number, it’s worth noting that an awful lot of us will spend at least sometime at the top, too…the share of Americans who will spend at least one year by age 60 in the top 20 percent (is 7 in 10)…and in the top 1 percent (is 1 in 10).” – July 24, 2015, New York Times (Badger and Ingraham 2015).

A long-standing literature has focused on income dynamics at the bottom of the distribution, namely mobility into and out of poverty (Bane and Ellwood 1986; Duncan 1984). In comparison, there are relatively few contributions on the dynamics of affluence. Recent research has documented widening economic inequality in the United States, and increasing concentration of advantage at the top of the income distribution (Piketty and Goldhammer 2014; Stiglitz 2012), but just how accessible are the top income/earnings
positions? Has access to these top positions changed over time, resulting in a more permeable or rigid class structure among the highest earners?

In this paper, we will focus on three main research objectives. First, we aim to produce new estimates of the probability of accessing the top positions in the income and earnings distributions. Only one study (referenced above in the NYT quote and discussed later in more detail) has produced similar estimates, which we believe to be biased because of methodological shortcomings and data limitations. Second, we aim to use administrative data to examine trends in access to the top of the income and earnings distribution. Finally, we aim to explain fluctuations in access to the top income and earnings positions. Our research will contribute to a fuller understanding of inequality by revealing important information about a surprisingly understudied group: the very rich.

Background and Research Questions

In the aftermath of the banking and housing busts that precipitated the Great Recession of the late 2000s, questions about the sustainability of an ever-increasing gap between rich and poor piqued the public’s interest in economic inequality. Talk of 1- and 99-percenter is now common in newspaper reports and political ads, demonstrating that the issue of inequality has worked its way into the mainstream media and political classes.

Increasing inequality in the U.S. has been accompanied by rising instability in earnings and income. Gottschalk and Moffitt (2009; 1994) show that instability in male earnings rose rapidly in the late 1970s and grew steadily in the 1980s. Instability in family income followed a similar pattern, but has continued to grow more sharply in the 1990s. However, an increase in income instability overall may mask differential patterns at different parts in the income distribution. For example, the transitory variance in income may have increased more so at the bottom of the distribution and less so at the top. Additionally, knowing about trends in individuals’ income volatility reveals little about how individuals fare relative to others in the distribution. Hence, just because income volatility is high or increasing does not mean that there has been any corresponding change in access to certain positions in the distribution.

Still, research on income dynamics can be very useful for understanding changes in the overall structure of income/earnings regimes. The vast majority of this research has focused on the lower end of the socioeconomic scale: A long-standing literature on poverty dynamics, beginning with Duncan (1984) and Bane and Ellwood (1986), examines mobility into and out of poverty.

Similar research that focuses on the top-end of the distribution is scant and as a result very little is known about the transience in income/earnings for individuals at the top of the income distribution. Carroll (2010) finds that spikes in income among top earners are relatively common, suggesting potential for high turnover at the very top of the income distribution. A few other studies have examined short-term persistence in the very top positions. Auten and Gee (2009) found that 60 percent of the top 1-percent of income earners dropped from the top 1 percent within ten years. More recently, Auten, Gee and
Turner (2013) found that persistence in the top 1 percent for five consecutive years fluctuated around 35 percent, edging slightly lower during intervals that included a recession. Collectively, these studies reinforce the impression that income volatility is high for those in the top income/earnings positions – yet these findings say nothing about the share of the population who ever succeed in accessing the top or about whether access to these top positions has changed over time. Our goal is to tell a story about the structure and rigidity of the upper end of the income/earnings distribution in the United States and how it has changed over time.

A recent study by Hirschl and Rank (2015) draws on data from the Panel Study of Income Dynamics (PSID) and documents surprisingly high turnover at the top of the income distribution. They estimate that 11 percent of individuals experience family income in the top 1 percent between 25 and 60 and report that few individuals persist at the top over many consecutive years. Perhaps equally surprising, they estimate that as many as 70% of individuals experienced family income in the top 20 percent. However, as our preliminary findings suggest, we believe their study suffers from several methodological and measurement shortcomings that await clarification.

Growing inequality in the United States has occurred against a backdrop of a multitude of economic, demographic, and social changes since the mid-twentieth century. Overall, a large decline in the share of manufacturing jobs and the rise in availability of lower-paying, less secure retail and service sector jobs occurred, shifting the economic realities for low-wage workers. Low-wage workers saw their wages stagnate and decline over time, lowering the value of work for younger cohorts. More women join the labor force today than did several years ago (Appelbaum, Berhnhardt and Murnane 2003).

Furthermore, changes in the structure of the American labor force were accompanied by larger demographic and societal shifts. As more and more women were entering the workforce, so too were more women and minorities completing their education and pursuing higher education, resulting in a change in the stock of human capital over time. The feminist movement of the 1970s changed more than just colleges and workplaces – it also coincided with massive changes to family structure and reproductive timing. The rise and leveling of divorce, the delay in age at marriage and fertility, reduced fertility, and increases in single-parent households were indicators of a nation-wide demographic shift away from the idyllic post-war era of the 1950s (DiPrete and Buchman 2013). Any one of these changes could have impacted the rigidity of the upper end of the income/earnings distribution by making it harder or easier for individuals to take up top positions. Our second goal is to determine how much of the changing rigidity in the income/earnings structure is due to changes in overall population composition.

Data and Measures

Research on the affluent is often hampered by myriad measurement challenges such as measurement error in income and earnings information, population coverage issues, and low statistical power. Administrative data provide a remedy for these issues but often are unavailable because of prohibitive access restrictions (Mervis 2014). Additionally,
research questions like ours call for the use of panel data (e.g. linked observations over time), adding another layer of difficulty in obtaining suitable data.

We will use three different data sources for our analysis: The Panel Study of Income Dynamics (PSID), the Survey of Income and Program Participation linked to lifetime earnings records from the Social Security Administration (SIPP-SSA), and the 1987 Family Panel collected by the Statistics of Income (SOI) Division of the United States Internal Revenue Service (IRS).

The PSID has collected income and earnings data from its respondents on an annual basis since 1968 and biennially since 1997. Besides its long observation window, one of the main advantages of the PSID is that it collects not only earnings information but also detailed information on total family income. This makes the PSID an attractive choice, but the data source is not without serious limitations.

First, the PSID is subject to unit non-response as well as item non-response on income and earnings information. Despite PSID’s efforts in accounting for the potential biasing effects of non-response through longitudinal weights and through imputation of income information, the administrative data used here will greatly reduce the potential bias introduced by non-response. Second, while the PSID income and earnings data track other government statistics well for most of the distribution, they may fail to adequately capture the top 5% and certainly fail to reliably capture the top 1% (Gouskova and Schoeni 1997). As Piketty and Saez (2003) state, “IRS tax registers provide the only reliable basis to assess trends in access to the top 1% of income earners”. Even if the Current Population Survey (CPS) did capture the top 1% adequately, as some have claimed (Burkhauser et al. 2012; but see Smeeding and Garfinkel 2011), it would still not provide the required longitudinal information for this study. Third, though the PSID has grown over time to cover nearly 25,000 individuals residing in more than 9,000 families in the latest available wave, the sample size within each year is small enough to prohibit a reliable assessment of trends for the richest 1% included in the survey. Since this group consists of just between 30 and 60 cases depending on the survey year, even a small amount of noise can be expected to produce a greatly inflated estimate of permeability (Ansolabehere et al. 2015).

The SIPP-SSA, in comparison, provides a much larger sample size and full earnings histories for a wider range of birth cohorts. SSA records provide W-2 earnings histories for over 700,000 respondents of the SIPP and of these about 340,000 have earnings records for an interval longer than 20 years. About 48,000 respondents have earnings records available for every year in the 34-year period from 1978-2011. Unfortunately, the SIPP-SSA data are limited to earnings and do not include data on family income. Still, this data source greatly alleviates concerns about measurement error and statistical power.

The 1987 ‘Family Panel’ was created by the Statistics of Income (SOI) Division of the United Stated Internal Revenue Service (IRS) and represents the cohort of families filing returns for tax year 1987 and every year since. In the initial year about 90,000 returns
were included in the Family Panel. According to Johnson, Moore and Schreiber (2009), “the SOI stratified sample design oversamples high-income taxpayers to ensure accurate estimates of the often unique financial characteristics of this elite group,” making it an ideal data source for this project. For more information about the 1987 Family Panel, see Nunns et al (2008).

The joint analysis of survey and administrative data provides us with at least two unique analytic opportunities: First, we can accurately estimate the proportion of individuals that ever experience affluence during adulthood based on fully observed lifetime incomes or earnings. Second, we can produce these estimates for multiple cohorts of individuals, allowing us to assess trends in affluence dynamics over time.

As the main outcomes in our analyses, we will measure whether an individual occupies a position in the top 20, 10, 5, and 1 percent of the income/earnings distribution in a given year. We will produce a set of measures based on family income from the PSID and IRS 1987 Family Panel and another set of measures based on individual earnings from the PSID and the SIPP-SSA. Unfortunately, family income is unavailable for the SIPP-SSA sample.

In the PSID, family income is measured by combining the pre-tax income for all family members from wages, pensions, property, capital gains, transfers and many government programs and adjusting for family size (done several ways as robustness checks). The adjusted family income and individual earnings for each respondent will then be compared to income/earnings distribution for a given year, which will be calculated from annual weighted cross-sections of the full PSID sample. From this information, we will create binary variables indicating whether or not an individual has income/earnings above the $p^{th}$ percentile in a given year.

In the SIPP-SSA, only individual earnings are available for analysis. Using a process similar to the one described above for the PSID, we will create a series of binary variables that indicate whether an individual has earnings above the $p^{th}$ percentile in a given year. As with the PSID, in each observed year individual respondents’ earnings will be compared to distributions drawn from annual cross-sections of the full SIPP-SSA sample.

For the IRS 1987 Family Panel (IRS-87FP), our measure will be based on family income for the ‘tax family,’ adjusted for family size. A ‘tax family’ consists of a taxpayer, spouse and all dependents as claimed on the Form 1040. The IRS-87FP concept of ‘family’ is slightly different than the typical ‘household’ concept employed by many other national surveys because “married couples who elect to file separately are treated as two distinct tax families. Only the partner whose return was selected into the sample was included in the panel. As a result, the tax family differs significantly from the more common household measure used by many national surveys” (Johnson, Moore and Schreiber 2009). However, the concept of family in the PSID is also different from other widely used data sources in that not everyone in the same household is counted as a part of the PSID family. In our case, the atypical ‘family’ concepts employed by both the IRS-87FP
and PSID facilitate more comparable estimates.

For all analyses, we consider adulthood to begin at age 25. The selection of this age is somewhat arbitrary but motivated by our desire to accommodate the lengthened transition to adulthood, which sees more individuals residing at home into their early twenties. By beginning the analyses at age 25 we are giving individuals time to complete education and transition into their adult living arrangements and careers. We test the sensitivity of our results to different lower age limits. Importantly, we also account for the household status of individual respondents (dependent or not), for instance to avoid miscounting parental income as the income of young adult children living at home.

Analytic Strategy

To begin, we will estimate cumulative probabilities for experiencing annual income/earnings at the top of the distribution, that is, the probability of joining the top 20%, top 10%, top 5%, and top 1% by certain ages and spanning individuals’ working lives for as long as the observational data will allow. For Aim 1, we will replicate the results by Hirschl and Rank (2015) using the same dataset and methodological approach they use: the PSID and a standard life table approach. We will, however, carefully consider the consequences of left censoring (i.e. whether the outcome was properly treated as a repeatable outcome) and the handling of adult dependents, none of which were addressed in this prior research.

The life table approach is commonly employed by demographers and biostatisticians to study mortality and other outcomes (see Box-Steffensmeier and Jones (2004)). The method involves calculating age-specific hazard estimates (i.e. the probability of experiencing an event, in this case income/earnings above a given percentile, conditional on never previously experiencing the event) and then adding these estimates to a life table that calculates the ‘prevalence’ of the outcome by certain ages. These cumulative probabilities are also known as Kaplan-Meier failure estimates, and can be expressed in the form:

\[
\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}
\]

where \(n_i\) is the number of individuals who have not yet experienced an event at the beginning of interval \(i\), \(d_i\) is the number of individuals who experience the event during the interval \(i\) and \(i+1\), and \(S(t)\) is then the percent ‘surviving’ to time \(t\). The failure is \(1-S(t)\), and indicates the percent of individuals who have experienced the event in question by interval \(t\).

The life table approach is a powerful tool when the analysis has the ability to observe an entire cohort from the time they enter the risk set until the time they are no longer at risk of experiencing the outcome – often that means a very long time. But because detailed panel data are not always available, the methods are often employed to instead produce synthetic estimates. The main difference between actual and synthetic estimates is that instead of following a single cohort over a long period of time, several different cohorts
all spanning different parts of the age spectrum are observed simultaneously. The age-specific probabilities from these differing cohorts are put together into one single life table to produce a set of life table estimates.

These estimates are most accurate in a stable population environment and are most problematic when younger cohorts are likely to differ greatly from older cohorts with respect to the outcome in question. Because later-age estimates of the age-specific hazards rely almost entirely on the experiences of older people in the synthetic cohort, aggregate estimates become biased toward the experiences of the older cohorts at an increasing rate (Grieger and Danziger 2011). In addition, this later-age bias can easily be exacerbated when applying these methods to estimate the hazard of experiencing repeatable events when there is uncertainty about whether or not the respondent has experienced the event before. This is because these methods were originally designed for outcomes like mortality where a researcher can be certain that an observed mortality “spell” is the first and only one that a subject will ever experience.

Next, we will extend the analysis from family income to individual earnings by replicating the analysis using the earnings-based measure in the PSID, which will facilitate a more direct comparison with replicates derived from the SIPP-SSA. Again, the purpose of utilizing the SIPP-SSA is to leverage the much larger sample size to assess the earnings dynamics at the very top of the distribution (1%).

The SIPP-SSA and IRS-87FP data will allow direct comparison of several cohorts to document secular trends in affluence dynamics. Table 1 summarizes each of the data sources and the respective coverage: The data allow for direct comparison of access to the top family income positions for 18 cohorts of adults to age 30 from 1987-2005 using the PSID and the IRS-87FP. They also enable direct comparison of access to the top individual earnings positions for 27 cohorts from 1979-2006 using the PSID and the SIPP-SSA.

Finally, we will compare access to top positions by older and younger cohorts to identify which key demographic covariates (race, gender, education, marital status, etc.) are associated with entry into the top income/earnings positions and to estimate how much of the difference between old and young cohorts can be explained by compositional changes in these covariates. We will rely on two approaches for examining the impact of demographic changes on our outcome. The first approach is via a regression framework, employing standard hazard/survival modeling techniques in order to estimate the hazard of accessing the top income/earnings positions as a function of key demographic covariates.

The second approach consists of a regression decomposition framework. For this part of the analysis, we will classify the respondents into two groups, representing a younger and an older cohort. Typical regression decomposition analyses rely on a switching framework, where identical regressions for each group are estimated and the coefficients switched one by one to estimate the average outcome for individuals of group A if they
had similar characteristics to individuals of group B. The typical regression decomposition can be expressed as:

\[
gap = \left( \overline{X}^1 - \overline{X}^2 \right) \hat{B}^1 + \left| \overline{X}^2 \left( \hat{B}^1 - \hat{B}^2 \right) \right|
\]

where \( \overline{X}^1 \) and \( \overline{X}^2 \) are row vectors consisting of the average values for each of the independent variables for group 1 and 2, respectively, and \( \hat{B}^1 \) and \( \hat{B}^2 \) are column vectors consisting of the coefficients from group-specific regressions. The first term represents the portion of the gap due to group differences in the distributions of the independent variables – or the difference due to coefficients. The second term represents the portion of the gap due to differences in unobserved endowments (Blinder 1973; Cain 1987; Fairlie 2005; Jones 1983; Oaxaca 1973).

Since the outcome we are interested in is binary, it is not ideal to use the regular Blinder-Oaxaca decomposition method because it would require the use of a linear probability model to obtain coefficient estimates. Rather, we are interested in a regression decomposition method suitable for a logistic regression model. Following Fairlie (2005), such a decomposition can be expressed as:

\[
gap = \left[ \sum_{i=1}^{n^1} \frac{F(X_iB^1)}{n^1} - \sum_{i=1}^{n^2} \frac{F(X_iB^2)}{n^2} \right] + \left[ \sum_{i=1}^{n^1} \frac{F(X_i^2B^1)}{n^2} - \sum_{i=1}^{n^2} \frac{F(X_i^2B^2)}{n^2} \right]
\]

where \( F(\cdot) \) is the logistic cumulative density function, \( X_i^k \) is a vector of independent variables for individual \( i \) in group \( k \), \( B^k \) is a vector of coefficients from a logistic regression of the outcome on the independent variables for group \( k \), and \( n^k \) is the number of observations in group \( k \). Essentially, the expression in the first set of brackets is the proportion of group 1 with the outcome of interest and the expression in the second set of brackets is the proportion of group 2 with the outcome of interest. When the two bracketed terms are combined, the middle expressions subtract away leaving only the gap between group 1 and 2.

Although the first term gives the amount of the gap between groups explained by all observed values, we are interested in how much each specific demographic factor contributes. To find the portion of the gap explained by a single independent variable \( x \), samples consisting of equal numbers for each group must be used. First, each observation is assigned a predicted probability, which is derived from the group-specific logistic regression. Then, the observations in each group are ordered and paired into \( n \) comparison pairs, with the observations having the lowest predicted probabilities from both groups compared to one another, the observations with the second lowest predicted probabilities compared to one another, and so on. The portion of the gap due to differences in the distribution of \( x_1 \) is:
where $\beta_j^p$ is a coefficient from a pooled logistic regression with both groups and $x_{ij}^k$ is the value of the $j^{th}$ independent variable for group $k$ in comparison pair $i$. Essentially, the amount of the gap due to a specific variable is the difference between estimates when the $x$'s for group one are switched to group two, holding all other variables constant. As such, the amount of the gap due to $x_2$ is:

$$\frac{1}{n} \sum_{i=1}^{n} [F(\hat{\beta}_0^p + x_{1i}^1 \hat{\beta}_1^p + x_{1i}^2 \hat{\beta}_2^p + x_{1i}^3 \hat{\beta}_3^p ... ) - F(\hat{\beta}_0^p + x_{1i}^2 \hat{\beta}_1^p + x_{1i}^2 \hat{\beta}_2^p + x_{1i}^3 \hat{\beta}_3^p ... )]$$

and the amount of the gap due to differences in the distribution of $x_3$ is:

$$\frac{1}{n} \sum_{i=1}^{n} [F(\hat{\beta}_0^p + x_{1i}^2 \hat{\beta}_1^p + x_{1i}^2 \hat{\beta}_2^p + x_{1i}^3 \hat{\beta}_3^p ... ) - F(\hat{\beta}_0^p + x_{1i}^3 \hat{\beta}_1^p + x_{1i}^3 \hat{\beta}_2^p + x_{1i}^3 \hat{\beta}_3^p ... )]$$

For our analysis, groups 1 and 2 represent the younger and older cohort, respectively. Each of the $x$'s represents one of our key demographic variables. Because the number of observations in groups 1 and 2 are not equal (thus disrupting a one-to-one match for the comparison groups), we will draw a random sample of observations from the larger group to match the number of observations in the smaller group. We conduct this randomization several times as prescribed by Fairlie (2005). Also, as the models are non-linear, the contribution of a single variable could depend on the order of the switching and we plan to re-estimate the model to assure our results are robust to the order in which the covariates were introduced.

Preliminary Findings:

We have begun to replicate the PSID analyses by Hirsch and Rank (2015) and have estimated the probability that a 25-year-old reaches income at or above the 80th, 90th, 95th, and 99th percentile by ages 35, 45, and 60 (see Figure 1). Unlike Hirschl and Rank, we restrict our analyses to actual rather than synthetic cohorts (individuals with complete observation windows). These preliminary analyses yield two important findings: First, the permeability of the top of the income distribution – i.e. access to the 80th, 90th, and 95th percentile – is generally decreasing over time.

This pattern of cohort change is problematic for estimations using synthetic cohorts, like Hirschl and Rank (2015), since it violates the underlying assumption that the full lifetime income and earnings profiles of younger individuals – which are, of course, unobservable – will resemble those of older cohort members. Second, the noise around the estimates
for the 99th percentile makes it difficult to draw any firm conclusions about the very top of the distribution. With only about 35-70 individuals in the top 1% of the PSID sample, even a small amount of noise is likely to introduce significant bias in the estimated probability of joining the very top of the distribution.

In summary, we find so far that the synthetic cohort approach hides large and substantively interesting variation across cohorts (as in Figure 1) and thereby misrepresents current levels of affluence dynamics.

Tables and Figures:

Figure 1: Probability a 25 year-old will attain family income at or above a given percentile.

Note: Analyses of PSID data by the authors. Reading example: The probability of attaining family income at or above the 90th percentile by age 35 was about 30% for an individual who was 25 years old in 1967 (the left-most point of the green line in the figure for the 90th percentile) and about 18% for an individual who was 25 years old in 2000 (right-most point of the same green line).
Table 1: Coverage of Data Sources

<table>
<thead>
<tr>
<th>Source:</th>
<th>Measures:</th>
<th>Coverage:</th>
<th>Number of Cohorts Observed Until:</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSID</td>
<td>Ind Earnings, Fam Income</td>
<td>1967-2012</td>
<td>Age 30  40  Age 35  35  Age 40  30  Age 45  25  Age 50  20  Age 55  15  Age 60  10</td>
</tr>
<tr>
<td>SIPP-SSA</td>
<td>Ind Earnings</td>
<td>1979-2011</td>
<td>Age 30  27  Age 35  22  Age 40  17  Age 45  12  Age 50  7  Age 55  2  Age 60  2</td>
</tr>
<tr>
<td>IRS-87FP</td>
<td>Fam Income</td>
<td>1987-2010</td>
<td>Age 30  18  Age 35  13  Age 40  8  Age 45  3  Age 50  3  Age 55  3  Age 60  3</td>
</tr>
</tbody>
</table>

Reading Example: The PSID is capable of producing estimates of access to top income positions from the onset of adulthood (age 25) to age 30 for forty cohorts beginning in 1967. The IRS-87FP is capable of producing estimates for eighteen cohorts beginning in 1987.

References:


