On the Nature of Entrepreneurship*

Anmol Bhandari
University of Minnesota

Tobey Kass
Office of Tax Analysis

Thomas J. May
California State University, Fullerton

Ellen R. McGrattan
University of Minnesota

Evan Schulz
Internal Revenue Service

August 6, 2025

ABSTRACT

This paper examines the nature of entrepreneurship using a novel panel dataset based on U.S. Internal Revenue Service data. We analyze income growth trajectories and the determinants of entrepreneurial choice for business owners. Compared to prior household survey-based studies, we find that self-employed individuals have significantly higher average incomes and steeper income growth than their wage-earning matched peers. Contrary to survey evidence, we find a limited role for non-pecuniary motives, uninsurable risk, and liquidity constraints driving entrepreneurial choice. Prior work experience predicts entrepreneurial entry. Entry and exit rates into self-employment remain stable over time, including during the Great Recession.

Keywords: Entrepreneurship, occupational choice, taxation
JEL classification: E13, H21

* Corresponding author: McGrattan, University of Minnesota, 1925 4th Street South, Minneapolis, MN 55455. An interactive companion to the paper, including access to the data, is available at www.econ.umn.edu/∼erm/support/OTNOELive.html. The authors thank Anne Parker and Barry Johnson for facilitating this project through the Joint Statistical Research Program of the Statistics of Income Division of the United States Internal Revenue Service. We acknowledge support from the National Science Foundation (Award #2214248). For their valuable feedback on earlier drafts, we thank the editor, four referees, and seminar participants at the Office of Tax Analysis, Federal Reserves, Stanford, Harvard, Yale, Princeton, ASU, NYU, and Toronto and conference participants at NBER, SED, and Hoover Institution. This research was conducted while May and McGrattan were IRS employees without pay under an agreement made possible by the Intragovernmental Personnel Act of 1970 (5 U.S.C. 3371-3376) and Kass was an employee at the U.S. Department of the Treasury. Any findings, interpretations, opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views or political positions of the IRS or the U.S. Department of the Treasury, or the NSF. All results have been reviewed to ensure that no confidential information is disclosed. All data work for this project involving confidential taxpayer information was done at IRS facilities, on IRS computers, by IRS and Department of the Treasury employees, and at no time was confidential taxpayer data ever outside of the IRS computing environment.

1 Introduction

In this paper, we use U.S. administrative tax data to assemble a novel longitudinal database of pass-through business owners—one that is suitable for analyzing patterns of income growth and determinants of entrepreneurial choice for a large population of self-employed individuals. Collectively, pass-through owners, including sole proprietors, partners, and S corporation owners, account for over 50 percent of all business net income in the United States and have been central to studies of entrepreneurship.[1] However, because their businesses are privately held, previous research has mostly relied on household survey evidence limited by small samples, short panels, and top-coded incomes.[2] As a result, much attention has been paid to the typical individual that chose self-employment rather than the typical dollar earned in self-employment. The typical self-employed individual is commonly portrayed as someone that faces inordinate risk and high startup costs and yet chooses self-employment at lower pay than what could be earned in paid-employment because running a business offers non-pecuniary benefits. Once we include all dollars earned in self-employment, we do not find strong evidence of inordinate risk, high startup costs, or non-pecuniary motives, and, on average, we find the self-employed have much steeper income growth profiles than paid-employed peers with the same characteristics.

To construct our longitudinal dataset, we utilize the Statistics of Income (SOI) Databank, which combines records from the Social Security Administration (SSA) and the Internal Revenue Service (IRS) and provides us with demographic information such as age, gender, marital status, and the number of children, as well as information on employment status, occupation, industry, own incomes, and household incomes. We use machine learning algorithms to impute additional information such as educational attainment and various measures of skill. Our measure of self-employment income is the sum of incomes from proprietorships (Form 1040, Schedule C net profits), partnerships (Form 1065, Schedule K-1 ordinary business income), S corporations (Form 1120-S, Schedule K-1 ordinary business income), and own-business compensation (Form W-2 wages from S corporations that they own). Paid-employment income is wage income (Form W-2 wages) less any own-business compensation. Along with the income measures, we use auxiliary data such as the number of employees and business gross profits to classify individual-year pairs as self-employed, paid-employed, or non-employed. We then create three datasets for the period 2000–2015: an unbalanced panel with all individuals aged 25 to 65; a balanced panel with individuals aged 25 to 65 born between 1950 and 1975 and alive in 2015; and a second balanced panel with the additional restriction that we have occupational information needed for classifying skills.

We compute cross-sectional moments for the unbalanced panel in order to check representation of the self-employed in the widely used U.S. Current Population Survey (CPS), and we construct life-cycle profiles for our main balanced panel in order to compare the income growth of self- and paid-employed individuals. In the IRS versus CPS comparison, we use the same criteria for categorizing the self-employed but find stark differences in median and mean incomes across all ages. In contrast, when we repeat the exercise for the paid-employed, we find that the median and mean incomes are close at all ages. We investigate the sources of differences for the self-employed by comparing observable characteristics and distributional statistics for the CPS and IRS samples. We then decompose the aggregate mean income differential for the self-employed into differences

[1] See the *Statistics of Income* Business Tax Statistics (2020), www.irs.gov/statistics.

[2] Prominent examples are Lazear and Moore (1984) with the Current Population Survey, Evans and Leighton (1989) with the National Longitudinal Survey of Youth, Hamilton (2000) with the Survey of Income and Program Participation, Hurst and Lusardi (2004) with the Panel Study of Income Dynamics, and Moskowitz and Vissing-Jorgensen (2002) and Kartashova (2014) with the Survey of Consumer Finances. For a comprehensive set of references, see Parker (2018).

in representation and differences in mean income across subgroups. When we group individuals by industry and compare IRS and CPS mean incomes for those earning above a certain income threshold, we find that most of the overall difference in the IRS-CPS gap is attributable to the right tail of the distribution, and a large fraction is accounted for by a few sectors such as professional services and health care.

In order to estimate life-cycle income profiles using our IRS balanced panel data, we exploit the presence of many cohorts to separately estimate age and time effects for disaggregated groups in the population. To implement the approach, we define groups as a Cartesian product of time-invariant characteristics, which depend on education levels, skills, primary industries, demographics, and employment histories. Once we have the group information, we estimate a flexible specification for individual incomes at a particular age with three components. The first component is an individual-level fixed effect meant to capture latent abilities, preferences, and other unobservable characteristics. The second component is a group-specific time effect that captures movements common to all individuals in the group, such as the effect of the Great Recession occurring in 2008–2009. The third component is an age effect that depends on the individual's cohort and group and is meant to capture changes in income over the life cycle. Our identification scheme assumes age effects are similar across binned cohorts and, with the long panel aspect of our dataset, allows us to estimate the time and age effects for all subgroups.

Using the estimates of time and age effects, we compare income profiles for self-employed groups to similar peers in the paid-employed groups and find patterns contrary to the previous survey evidence of Hamilton (2000) and Hurst and Pugsley (2011). Of particular interest are comparisons of individuals within a group we call *primarily employed*, who have at least 12 of the 16 sample years in self- or paid-employment with at most one gap year of non-employment. Our headline comparisons within this group of primarily employed individuals are the *primarily paid-employed* who have at least 12 years in paid-employment with at most one gap year of non-employment and all others who—in our terminology—have *tried self-employment*. The latter group includes business owners that do little or no paid-employment work and individuals that have experience in both self- and paid-employment and possibly switch back and forth; collectively, they earn virtually all self-employment income. When comparing income profiles, we compute average incomes at age 25 and add group-level age and time effects.

These estimated income profiles are starkly different for our aggregate self- and paid-employed groups. Average incomes at age 25 are roughly the same, but the estimated income profiles are much steeper for those who have tried self-employment when compared to their paid-employed peers. By age 55, the estimated income is $134 thousand for the self-employed and $79 thousand for the paid-employed. We also find stark differences when comparing the self- and paid-employed in various subgroups of industries, skills, and demographics.

Given the is considerable heterogeneity in the population of self-employed, we investigate the characteristics of a subgroup congruous with the common view that non-pecuniary benefits are an important driver of entrepreneurship. More specifically, we choose individuals from our primarily-employed group that have at least 12 years in self-employment and earn less on average over the sample years than primarily paid-employed peers with the same gender, education, skills, and so on. We find that the self-employed earning less than paid-employed peers are larger in number than those earning more—roughly 57 percent—but earn only 16 percent of the total income, which suggests that non-pecuniary motives could be guiding the occupational choice. If we repeat the exercise but include estimates of paid-employed benefits and self-employed income misreporting, we can affect the size of the group that would fit the narrative, but not the main finding that this group accounts for a small share of the total income earned by the self-employed.

We next ask whether the steeper income profiles in self-employment that we find compen-

sate for the additional risk inherent in entrepreneurship. We study idiosyncratic risk by comparing patterns of variability and persistence of income growth over the life cycle for both the self- and paid-employed. Relative to paid-employed, self-employed income changes exhibit greater dispersion—with the 10^{th} to 90^{th} percentile range roughly 2.5 times larger—and are more right skewed, where the difference in the Kelly skewness is about 0.1. To assess whether this risk is compensated, we parameterize a standard model of consumption risk—modified with a lower bound on consumption growth to allow for external sources of insurance—and compute an indifference curve for choosing self- versus paid-employment when varying the parameters of risk aversion and floors for consumption growth. Using estimates of risk aversion from the household finance literature, we find that it is easy to rationalize the patterns in our data if one is willing to accept that individuals are insured against the most adverse shocks. In other words, if potentially large upside gains and insured downside losses are associated with self-employment, then it is an attractive option.

We study aggregate risk by tracking income changes and exit rates of the self-employed and paid-employed during the Great Recession of 2008–2009. Not surprisingly, our estimated time effects in income growth show dramatic declines in sectors such as real estate and construction. Once we disaggregate by employment status, we find that most of the declines are attributed to self-employed subgroups in these cyclically sensitive sectors, with their paid-employed peers experiencing much more modest declines. Despite the large declines in incomes for the self-employed, we find little change in exit rates—in the aggregate or in cyclically sensitive sectors—during our sample that includes the Great Recession.

As in the case of exit rates, we find a remarkably constant entry rate into self-employment and no decline in the share of entrepreneurs in the population over our sample period, even during the Great Recession. Although this evidence might suggest that business dynamism is not in decline and that impediments to entrepreneurship are not large, we report additional evidence that speaks more directly to the question of whether entrants have sufficient liquidity, experience, and insurance. First, we follow the literature and test the hypothesis that entry rates are higher for homeowners experiencing an appreciation in house prices, which would indicate that collateral constraints are binding. We do not find evidence that they are. Second, we compare past asset incomes—interest, dividends, and capital gains—of current entrants and future entrants with the same characteristics to determine if they have more liquid resources to cover startup costs. We find that most entrants have lower asset incomes and less liquid wealth than those switching later. Third, we compare past labor income of current entrants and future entrants with the same characteristics to determine if they are successful in paid-employment before switching or unsuccessful and thus forced into self-employment. We find that most entrants have higher past labor income, which is indicative that they come in with on-the-job experience and that self-employment is not a backup option for low-paid workers. Finally, we analyze individual and business tax filings of business founders in the first three years of operation and compare the timing of positive incomes for both sets of filings. We find that almost all founders in our sample have positive income on their individual tax form in the first year of operation even though most have negative business net incomes and no external debt financing.

2 Data

In this section, we describe our main sample drawn from U.S. administrative tax records.[3] We start with details of our data sources and the definitions of self- and paid-employment income. We then describe algorithms to impute skill and education levels.

[3]Replication codes and detailed documentation are available at the IRS.

2.1 Sample

When constructing our sample, we start with records in the SOI Databank, which is a de-identified balanced panel of all living individuals with a U.S. Social Security number over the period 1996 to 2015. (See Chetty et al. (2018) for full details on this database.) For each individual there are rows—one for each year—and columns recording demographic information from the SSA (such as age and gender) and economic data from tax filings (such as information on individual income tax forms and attachments).

The SOI Databank includes household-level income reported on Form 1040.[4] We merge in individual-level information on wages and salaries reported to the IRS on Form W-2 for employees. For sole proprietors, we assign income from Schedule C separately by Social Security number. For owners with pass-through businesses—partnerships and S corporations—we merge in information from Schedule K-1 filings attached to Form 1065 and 1120-S, respectively.[5] The Schedule K-1 data are available beginning in 2000, and thus our sample period ranges from 2000 to 2015.

A summary of our IRS sample selection is shown in Table 1. We start with individuals in the SOI Databank between the ages of 25 and 65. For the sample period 2000–2015, there are 3.2 billion total person-year observations. We use this unbalanced panel of 25- to 65-year-olds when comparing IRS data to household survey data. To construct income profiles by age, we use records for all individuals born between 1950 and 1975 that are still alive in 2015. This balanced panel includes roughly 2.0 billion person-year observations. Because education and skill play an important role in income determination, we use occupational information for individuals in the balanced panel sample to impute skills. This data restriction narrows our sample to roughly 1.3 billion person-year observations.[6]

2.2 Income Measures

For each individual-year observation, we compute two sources of income. The first is a measure of *self-employment income* and is defined as the sum of net profit or loss from sole proprietorships (Form 1040, Schedule C, line 31), the individual's share of ordinary business income from partnerships (Form 1065, Schedule K-1, Part III, line 1), the individual's share of ordinary business income from S corporations (Form 1120-S, Schedule K-1, Part III, line 1), and finally the individual's wages (Form W-2, Box 1) paid by the S corporations they own.[7] The second is a measure of *paid-employment income* and defined as the wages (on Form W-2, Box 1) paid by businesses that are not owned by the wage earner. We refer to the sum of self- and paid-employment income as *total income*, although it does not include other categories of adjusted gross income on the tax forms. These measures are computed before tax and transfers, exclude most employer fringe benefits, and are deflated by the Bureau of Economic Analysis's (BEA) personal consumption expenditure price index and reported in thousands of 2012 U.S. dollars. Except where noted, adjustments are not made to account for potential income underreporting.

[4] We exclude from our baseline sample any individuals that exclusively use the simpler Forms 1040A or 1040EZ for our sample period because business owners must file the standard form. Including these individuals would lower estimates of average income from paid-employment and strengthen our main findings.

[5] Business net incomes of Subchapter C Corporate shareholders are not passed through to individual income tax forms until the companies distribute dividends or capital gains.

[6] Details of the imputations are provided in Section 2.3.

[7] Here, we omit capital gains as a source of self-employment income. Conceptually, at least a part of these gains reflects entrepreneurial investment (see Bhandari and McGrattan (2021) and Bhandari, Martellini, and McGrattan (2024)) and should be included with self-employment income, but systematically reclassifying these gains with administrative data is beyond the scope of this paper. Including such gains would strengthen our main findings.

Table 1: IRS Sample Selection

Sample	Description	Observations
	All Individuals	
(A)	Aged 25 to 65	3.2 billion
(B)	In sample (A) and born in years 1950–1975 & alive in 2015	2.0 billion
(C)	In sample (B) and has occupational information	1.3 billion
	Self-employed Individuals ($\lvert y^{\text{SE}}\rvert > 5,000$)	
(1)	In sample (A) and SE income criteria met	169.2 million
(2)	In sample (C) and SE income criteria met	97.7 million
(3)	In sample (C) and any of three SE criteria met	107.6 million
	– All three SE criteria met	29.3 million
	– SE income and gross profit criteria met	64.6 million
	– All other cases	13.7 million

Notes. Data for sample (A) include all living individuals aged 25–65 with a U.S. Social Security number over the sample period 2000–2015. Samples (B) and (C) are balanced panels. Income from self-employment is denoted y^{SE}. For individuals to be included in one of the self-employed samples, the absolute value of the self-employment income, $|y^{\text{SE}}|$, must exceed \$5,000 in 2012 dollars. Three additional criteria are checked. If $|y^{\text{SE}}|$ is greater than income from paid-employment, then the *SE income criteria* is met. If gross profits are greater than income from paid-employment, then the *SE gross profit criteria* is met. If the individual has an ownership share times the number of employees equal to 1 or more, then the *SE employee criteria* is met. The table reports the total number of observations in each sample.

Although individuals can have both paid- and self-employment income, we assign individuals to distinct employment categories each year based on a test designed to gauge their primary activity. We use a two-step procedure. First, we classify an individual-year observation as self-employed (SE) or not using multiple criteria designed to capture active business ownership. Second, we classify the remaining observations as paid-employed (PE) if their income from paid-employment exceeds a threshold and non-employed (NE) otherwise.

Definition 1. An individual-year pair is classified as *self-employed* (SE) if the absolute value of self-employment income exceeds \$5,000 (in 2012 dollars) and at least one of the following criteria is met: (*i*) *income criteria*: the absolute value of their self-employment income is greater than their paid-employment income; (*ii*) *employee criteria*: the sum across businesses of the individual's ownership share times the number of its employees is larger than 1; or (*iii*) *gross profit criteria*: the sum across businesses of the individual's ownership share of gross profits (receipts less cost of goods sold) is in excess of the individual's paid-employment income.

We take the absolute value of the self-employment income because young entrepreneurs incur significant expenses when building up their businesses, and many have losses. The second additional criterion is added because hiring employees is indicative of owner attachment to self-employment. The third criterion allows for the fact that many successful business owners pay themselves little

income to minimize taxes but earn incomes later when selling their businesses. In Table 1, we report counts for the alternative samples and definitions of the self-employed. Using the full population of 25- to 65-year-olds, noted as "sample (A)," we have 169 million observations that meet criteria (i). We denote this group as "IRS self-employed sample (1)" and use it later in our comparison of IRS and CPS data. Using the baseline sample with a balanced panel and occupational information, noted as "sample (C)," we have 98 million observations that meet criteria (i). We denote this group as "IRS self-employed sample (2)" and use it later to compare our unbalanced and balanced panels. If we loosen the restriction and also include business owners that meet at least one of the three criteria listed above, our sample of self-employed has 108 million observations. We denote this group as "IRS self-employed sample (3)," which is the baseline sample for our longitudinal study. Note that this sample is not much larger than sample (2) because most owners have more income in self-employment than in paid-employment. If we break it down further, we find that 29 million owners meet all three additional criteria, and 65 million owners meet the two additional criteria but do not have employees.

Our notion of self-employment is intentionally distinct from papers such as Smith et al. (2019), DeBacker, Panousi, and Ramnath (2022), Garin, Jackson, and Koustas (2022), Abraham et al. (2020), and Lim et al. (2019), who all use IRS data to study business incomes. Smith et al. (2019) classify all individual recipients of Schedule K-1 as self-employed. Our definition excludes 43 million of the 138 million individual-year K-1 recipients in our sample from being classified as self-employed. These are cases in which an individual receives little income from business filings. While this is not a concern for studies of top incomes, which is the focus of Smith et al. (2019), our focus is to learn about returns to entrepreneurship. Therefore, we deliberately use a more conservative test when categorizing entrepreneurial activity. DeBacker, Panousi, and Ramnath (2022) use a panel that tracks tax filers for up to 32 years using the SOI sample from 1987. While this has the benefit of being a long panel, the number of self-employed individuals that are studied shrinks to roughly 2,000 observations over a few cohorts. Such a restrictive sample would be unsuitable for achieving our two main goals, which are (i) calculating life-cycle income profiles using overlapping cohorts to infer time and age effects and (ii) understanding the determinants of self-employment by comparing outcomes for narrowly defined groups—some of whom enter self-employment and some of whom do not. Garin, Jackson, and Koustas (2022) and Abraham et al. (2020) focus on Schedule SE filers. This focus is not suitable for our analysis because it misses a significant fraction of business owners, namely, entrepreneurs who make losses and S corporation owners that do not file Schedule SE. Lim et al. (2019) focus on independent contractors that receive a Form 1099 and have less than $10,000 in deductions, excluding car and travel expenses. While these individuals are included in our sample, restricting the analysis to this group would eliminate a significant fraction of self-employment income.

Next we define paid- and non-employed categories.

Definition 2. An individual-year pair is categorized as *paid-employed* (PE) if it is not already categorized as self-employed and if the paid-employment income of the individual in that year exceeds $5,000 (in 2012 dollars).

Definition 3. An individual-year pair is categorized as *non-employed* (NE) if it is not already categorized as SE or PE.

A potential concern is whether we classified individuals as non-employed when they were actually employed but missing in the SOI Databank. To partly address this concern, we cross-check the relevant items on Form 1040 with the individual's Form W-2, Schedule C, or Schedule K-1 if

available.[8]

2.3 Imputations for Skills and Education

A large empirical labor literature focuses on skills and education as determinants of income. In this section, we use data from tax filings—most notably, information on occupations—and data-trained classification algorithms to impute indicators of skill and education. We later use the estimates when analyzing characteristics of subgroups of tax filers.

2.3.1 Skills

After signing and dating the tax form, individual tax filers and their spouses are asked to self-report their occupation, which is summarized in the IRS data as a character string. The occupational information is available for electronically filed (or *e-filed*) returns for tax years 2005 and later, with the exception of 2012. For the sample of individuals born between 1950 and 1975, 89 million individuals e-filed at least once in the years that these occupation strings are available. We are able to assign skill values to the subset of 80 million individuals in our main sample.

First, 73 million individuals provide usable occupations, which can be mapped directly to a standard occupational classification (SOC) code.[9] For these individuals, we assign skill values using the procedure of Lise and Postel-Vinay (2020). The idea is to create a mapping between the SOC codes assigned to individuals and their cognitive, interpersonal, and manual abilities. This is done with the aid of the Occupational Information Network (O*NET) summary of skill requirements needed for each occupation. Since the summary of requirements is long for each occupation, Lise and Postel-Vinay (2020) use a principal component analysis (PCA) to construct indices—keeping the top three (orthonormal) components and ensuring that occupations requiring mathematics are encoded as "cognitive," occupations requiring social perceptiveness are encoded as "interpersonal," and occupations requiring mechanical knowledge are encoded as "manual."

Second, there are 7 million individuals for whom we impute a skill value.[10] For these individuals, we apply a k-nearest neighbor classifier for the imputation using information on k "neighbors" from the subsample of the 73 million individuals that have a valid SOC code and assigned skill values. The neighbors share the same gender, marital status, birth cohort, and two-digit NAICS industry code and are nearest in the paths of employment status and incomes.[11] For each subgroup, we operationalize choosing near neighbors in the case of time-varying income variables by applying a PCA that maps a high-dimensional vector of statistics from our data to a lower-dimensional vector of moments. Inputs to the PCA are paid- and self-employment income in each sample year and moments of total income averaged across sample years. The specific moments are the mean, the standard deviation, the minimum, and the maximum, with the latter three normalized by dividing by the mean. The number of principal components depends on our choice of the fraction of variance to be explained, which we denote here by v. Thus, we have two parameters to choose: the number of neighbors k and the fraction of variance v—and we assume they are fixed for all subgroups.

We choose parameters to maximize the predictive accuracy of the k-nearest neighbor classifier. To do this, we pull a random sample of subgroups and split them into three subsamples: 70 percent for training, 20 percent for tuning, and 10 percent for validation. For each (k, v) pair, we use the

[8] For instance, suppose the wage reported on Form W-2 for an unmarried filer is zero, but the wage reported on Form 1040 is positive. In this case, we use the Form 1040 wage data.

[9] We thank Raj Chetty and his team for providing us with a mapping between the strings and the SOC codes.

[10] For instance, business owners might fill in "self-employed," which is not a valid SOC code.

[11] In Section 4.2, we group individuals into four different categories of employment status based on attachment to and type of work.

Table 2: Summary Statistics for Main IRS Sample

Statistic	Total Sample	Self-Employed	Paid-Employed	Non-Employed
Observations (Mil.)	1279.9	107.6	940.9	231.4
Shares (%)				
Counts	100	8.4	73.5	18.1
Total income	100	15.2	84.6	0.2
SE income	100	97.4	2.4	0.2
PE income	100	2.7	97.1	0.2
Incomes (2012$, Thous.)				
Mean, total income	49.0	88.6	56.4	0.6
Percentiles, 10^{th}	0.0	5.9	14.0	0.0
25^{th}	11.4	12.7	24.8	0.0
50^{th}	32.8	29.0	41.4	0.0
75^{th}	58.6	78.4	65.3	0.3
90^{th}	95.3	193.3	100.1	3.0
Mean, SE income	6.5	75.0	0.2	0.1
Percentiles, 10^{th}	0.0	−6.0	0.0	0.0
25^{th}	0.0	10.2	0.0	0.0
50^{th}	0.0	23.1	0.0	0.0
75^{th}	0.0	64.1	0.0	0.0
90^{th}	2.9	167.8	0.0	0.0
Mean, PE income	42.6	13.6	56.2	0.5
Percentiles, 10^{th}	0.0	0.0	14.0	0.0
25^{th}	5.2	0.0	24.8	0.0
50^{th}	30.1	0.0	41.4	0.0
75^{th}	55.3	1.4	65.2	0.0
90^{th}	88.2	29.2	99.6	2.6
Education and skills (%)				
College educated	52.8	56.5	56.4	36.7
Cognitive	52.4	59.0	55.3	37.8
Interpersonal	58.6	56.8	62.5	43.7
Manual	37.6	39.2	37.1	39.1
Primary industry (%)				
Agriculture	0.7	1.4	0.8	–
Mining	0.3	0.4	0.3	–
Utilities	0.1	0.1	0.2	–

See notes at end of table.

Table 2: Summary Statistics for Main IRS Sample (cont.)

Statistic	Total Sample	Self-Employed	Paid-Employed	Non-Employed
Primary industry (%)				
Construction	4.7	15.1	4.7	–
Manufacturing	7.7	2.8	10.2	–
Wholesale trade	2.5	2.6	3.1	–
Retail trade	5.3	7.5	6.4	–
Transportation	2.1	5.5	2.3	–
Information	1.0	1.1	1.3	–
Finance	2.1	3.2	2.5	–
Real estate	1.5	4.9	1.5	–
Professional services	6.0	13.4	6.6	–
Management	0.4	0.1	0.6	–
Administration	2.9	4.9	3.4	–
Education	0.3	0.6	0.3	–
Health care	4.0	8.3	4.5	–
Arts	0.6	2.1	0.6	–
Accommodation	2.4	3.7	2.9	–
Other services	2.4	11.5	1.9	–
Other NAICS	9.1	3.0	12.1	–
Missing NAICS	43.7	7.8	34.0	100
Employees and Profits				
Has employees (%)	3.2	32.9	0.4	0.7
Gross profits (2012$, Thous.)	22.2	249.0	1.1	2.5
Demographics				
Male (%)	50.2	73.0	50.8	37.0
Married (%)	61.2	67.2	62.5	53.0
Has children (%)	50.6	58.9	53.2	36.1
Mean number of children	1.0	1.1	1.0	0.7
Median birth year	1963	1962	1963	1964
Other incomes (2012$, Thous.)				
Mean, spousal wages	26.7	21.7	25.2	35.0
Mean, asset income	8.8	37.5	5.2	9.8
Mean, UI income	0.4	0.2	0.4	0.6

Notes: SE=self-employed and PE=paid-employed. These summary statistics are computed using individual-year observations in the IRS sample (C) in Table 1, with the criteria for self-employed in (3). See Section 2 for more details on the sample construction and income measures. Incomes and gross profits are reported in thousands of 2012 dollars. To ensure that no confidential information is disclosed, reported percentiles here and elsewhere are computed as an average of observations around the value listed in the table. Industries are classified by two-digit NAICS.

training data to train the classifier and make predictions for the tuning set. We use the validation data to test predictions out of sample. The result of this exercise is $k = 11$ and $v = 75$ percent. With these parameters, we apply the classifier to impute skill values for 7 million individuals without usable SOC codes.

2.3.2 Education

The only indicators of education in the IRS microdata are occupation strings with "student" and tuition payment statements (Form 1098-T) filed by eligible educational institutions starting in 1998. Relying on these indicators alone would miss a large fraction of the population. To ensure full coverage, we use a classification algorithm and source data from the Annual Social and Economic Supplement of the CPS to predict the likelihood of college attainment.

We define individuals as being "college educated" if they have completed at least an associate's degree—which would thus include bachelor's, master's, professional school, and doctorate degrees. All others are considered "not college educated." For each year t, we run the regression

$$\Pr(E_{it} = 1|X_{it}) = \text{CDF}(\beta_t X_{it}), \tag{1}$$

where $E_{it} = 1$ if the individual is college educated and 0 otherwise for t between 1995 and 2020. The function CDF in (1) is the cumulative distribution function of the standard normal, and variables included in X_{it} are as follows: gender; annual pre-tax wages and salaries; positive business income (equal to 0 if income is negative); negative business income (equal to 0 if income is positive); marital status; number of children (with separate variables for none, one child, and so on, up to nine or more); five-year birth cohort; SOC minor occupation code; and two-digit NAICS industry code.[12] When we used 80 percent of our CPS sample each year to train the classifier and 20 percent to validate the predictions, we were able to correctly predict the education level with 75 to 80 percent accuracy. Coefficients from the CPS-trained classifiers are used with microdata from the IRS to impute an education indicator for all tax filers in our sample of 25- to 65-year-olds.[13]

2.4 Descriptive Statistics

Table 2 provides summary statistics for the baseline sample. The first column corresponds to sample (C) in Table 1 and includes 1.3 billion person-year observations (or, equivalently, 80 million persons over the 16-year sample).[14] The remaining columns of Table 2 summarize information for those categorized as self-, paid-, and non-employed who respectively represent 8 percent, 74 percent, and 18 percent of the counts.

Total income defined as combined self- plus paid-employment income averages $49 thousand (in 2012 dollars), with a range across the distribution from zero at the 10^{th} percentile to $95 thousand at the 90^{th} percentile. Of the total income, individuals earn about $6 thousand in self-employment and around $43 thousand in paid-employment. Thus, the share of self-employment income in our

[12] Some IRS tax filers do not have a valid NAICS code and do not have a SOC minor code. Additional regressions were run using (*i*) the SOC minor codes with no NAICS; (*ii*) the SOC major code and NAICS; and (*iii*) NAICS but no SOC.

[13] All variables in X_{it} are available in the IRS data, although the IRS occupation field is only available for tax years after 2005 and later (not including 2012).

[14] In the online appendix, we report the same statistics as in Table 2, but for different samples. Our samples are designed to test the robustness of how we define self-employment and to provide a more detailed analysis of relevant subgroups. With that in mind, we produce Table 2 statistics for self-employed individuals restricting exclusively to the income criteria in Definition 1, and then separately for subgroups defined by gender, education, and whether their business is an employer.

sample is equal to 13 percent in our sample. In terms of average incomes, those categorized as self-employed earn roughly $89 thousand as compared to $56 thousand for paid-employed, with most of the income coming from self-employment. Comparing income distributions, we find more substantial right skewness for the self-employed when compared to the paid-employed.

Overall, education and skill levels are similar for self- and paid-employed, but there are notable differences in their primary industries. When sorting individuals into industries, we use specific criteria based on their employment status. For paid-employed individuals, the assignment is based on the two-digit NAICS code of the employer who pays them the highest W-2 wage. For self-employed individuals, the assignment is based on the business that generates the highest gross profits. We find that more than half of the self-employed population is in construction, professional services, health care, and other services, which is a considerably higher fraction than for the paid-employed.

Two of our criteria for categorizing individuals as self-employed require us to measure the number of employees and gross profits (defined as sales minus cost of goods sold). Not surprisingly, we find that the share of employers and the average gross profits from business are negligible for individuals categorized as paid- or non-employed, even though it is possible that they earn some self-employment income. For those categorized as self-employed, the average gross profit is $249 thousand, and roughly 33 percent have employees.

In terms of demographics, we find a higher majority of the self-employed population are male and married with children (listed as non-spousal dependents aged 25 or younger). Finally, in terms of other incomes, the self-employed file tax returns with lower average spousal incomes and higher average asset incomes compared to other groups.

3 Comparison to Current Population Survey

In this section, we compare cross-sectional moments for the IRS samples described in Section 2 to survey data from the CPS and show that there are stark differences between the respective samples of self-employed individuals, but not between the paid-employed.[15] In order to track down the sources of these discrepancies, we compare distributional information in the micro datasets and find the CPS is not representative for high earners in key industries for the self-employed.

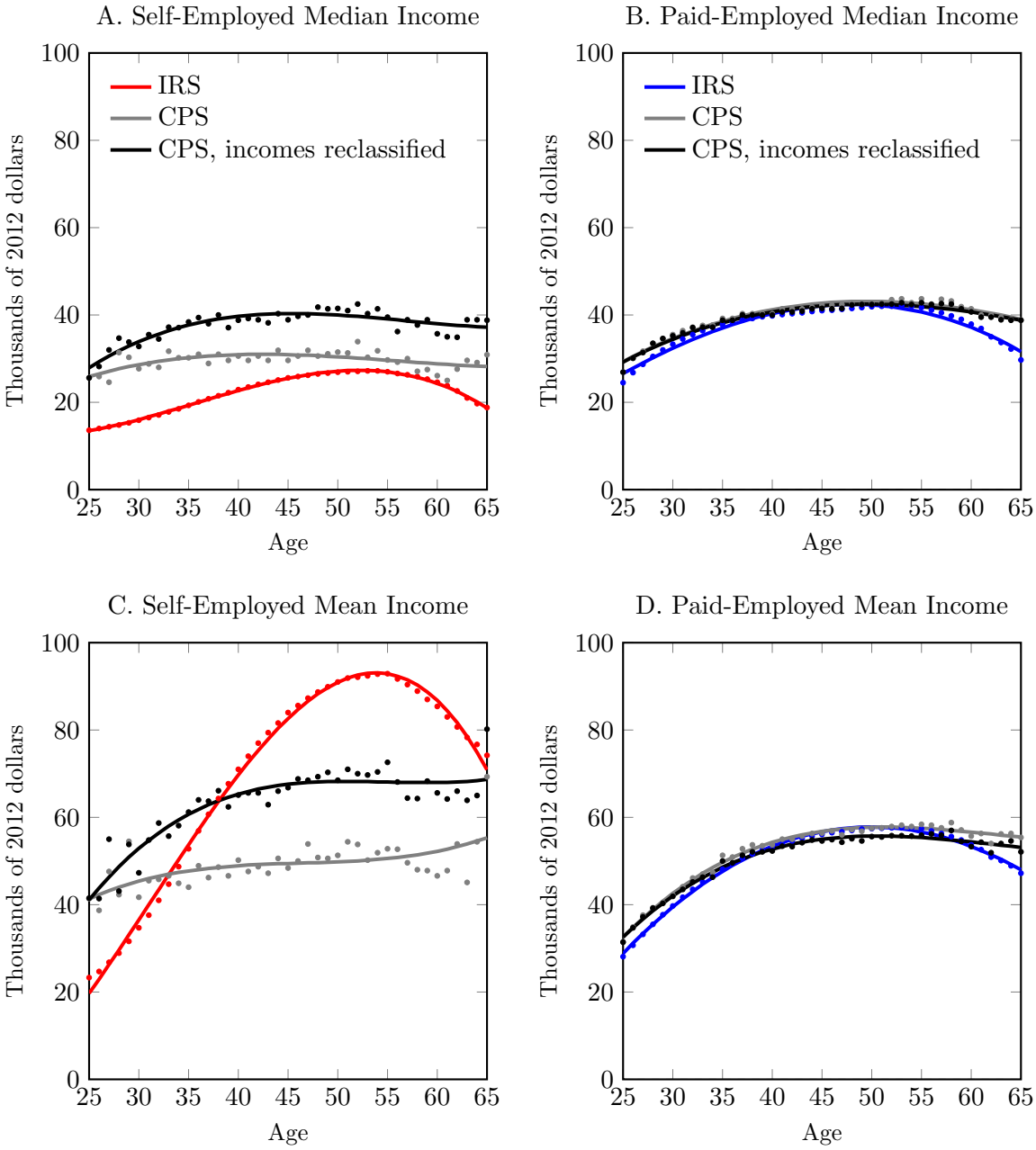
When reporting results for the CPS, we use two different classifications of self-employment income because the CPS treats incorporated business owners as wage and salary workers who are employees of their own business. The first classification is based on the self-employment income from a business or farm (IPUMS variable incbus) reported by the surveyed individual. With incorporated owners excluded, self-employed counts for the CPS should be lower than for the IRS. The second classification is based on the class of worker (IPUMS variable classwly) reported by the surveyed individual. If an individual reports that they run an incorporated business, we reclassify all wage income as self-employment income and assume they have no income from paid-employment. If an individual reports that they run an unincorporated business, we use their reported incomes from self- and paid-employment as we did for the first classification. Recall that our IRS sample only has pass-through owners. Thus, with incorporated owners included, self-employed counts for the CPS should be higher than for the IRS if owners of C corporations are included in the survey.[16]

To ensure consistency between the IRS and CPS, we apply the same procedure to categorize individual-year observations to self-, paid-, or non-employment. Since we do not have information

[15] The microdata are available at IPUMS CPS, University of Minnesota, www.ipums.org, with source data provided by the U.S. Census Bureau and Bureau of Labor Statistics.

[16] Levine and Rubinstein (2017) argue for including incorporated owners in studies of entrepreneurship.

Figure 1: Empirical Moments, IRS versus CPS



Notes: The IRS data are based on sample (A) of Table 1. Individuals are self-employed in a particular year and age if the absolute value of income from business exceeds $5,000 in 2012 dollars and the income from paid-employment. If these criteria are not met, but income from non-business wages and salaries exceeds $5,000 in 2012 dollars, then they are paid-employed. The first CPS sample uses reported incomes when categorizing the self- and paid-employed. The second CPS sample marked "incomes reclassified" reassigns all wage and salary income of incorporated business owners to self-employment before categorizing the self- and paid-employed. Medians and means of total income from both self- and paid-employment are shown in the plots. To ensure that no confidential information is disclosed, reported IRS medians are computed as an average of observations around the value listed in the figure.

about the employees or gross profits of business owners, we call an individual "self-employed" in a particular year if the absolute value of self-employment income exceeds $5,000 in 2012 U.S. dollars and exceeds the income earned from paid-employment. This definition for the self-employed is consistent with IRS sample (1) in Table 1. If these criteria are not satisfied, and income from paid-employment exceeds $5,000 in 2012 U.S. dollars, we call the individual paid-employed in the year. Otherwise, they are non-employed.

Finding 1. IRS data show lower median incomes for the self-employed at all ages as compared to the CPS, regardless of how earnings of incorporated owners are classified.

Since survey data have issues related to top-coding and small samples, most research on entrepreneurship has focused on the median incomes of the self-employed. In Panel A of Figure 1, we plot median total income—from self- and paid-employment—by age for those classified as self-employed in each dataset. The IRS medians (red dots) are based on sample (1) in Table 1 and computed as an average around the 50^{th} percentile to ensure no confidential information is disclosed.[17] Two samples are plotted for the CPS, namely, the gray dots representing median incomes without any reclassification of incorporated owner wages and the black dots with wages reclassified. For all samples, we include a polynomial fit to the data.[18] The figure shows that the IRS age profile for median income is below the CPS profiles, with the differences largest when we consider younger owners and those with incorporated businesses.[19]

To highlight the magnitudes of the IRS-CPS differences for the self-employed, we include the same comparison for the paid-employed in Panel B of Figure 1. As the figure shows, whether we classify incorporated owners as self-employed or paid-employed has no effect on the CPS moments in Panel B. More importantly, comparing either sample to the IRS, we find only a small gap between CPS and IRS median incomes—primarily for the youngest and oldest age groups. (Here and throughout, we use blue lines for data on paid-employed in the IRS sample.) The comparison with Panel B also provides a check on survey evidence that finds a self-employment "discount" if large non-pecuniary benefits are associated with being one's own boss and having flexible jobs (see, for example, Hamilton (2000), Hurst and Pugsley (2011), and Catherine (2022)). The lower median incomes for self-employed in the IRS sample and the CPS sample without incorporated owners are both consistent with this view.

Finding 2. IRS data show higher mean incomes for the self-employed at older ages as compared to the CPS, regardless of how earnings of incorporated owners are classified, and lower mean incomes at younger ages.

In the lower panels of Figure 1, we plot cross-sectional means of total income for each age. Panel C has results for the self-employed, and Panel D has analogous results for the paid-employed. In the case of self-employed individuals, the differences are large across the entire age profile regardless of which CPS sample is used in the comparison. The mean income for the self-employed in our IRS sample is $23 thousand at age 25 and rises to $93 thousand by age 55. The mean income for the self-employed in the CPS sample without incorporated owners reclassified is $41 thousand at age 25 and rises to only $53 thousand by age 55. If we add income for incorporated owners, we find a

[17] We exclude the top and bottom 0.01 percent outliers. Including these individuals adds more noise to the series of cross-sectional means but barely changes the polynomial fit.

[18] In 2010, the Census Bureau adopted a new method to avoid disclosure of top incomes in the CPS. Our quantitative results do not change if we use the post-2010 subsample.

[19] These findings are in line with Levine and Rubinstein (2017), who show that when compared to the paid-employed, median incomes are only marginally higher (about $5,000) for the incorporated self-employed using the National Longitudinal Survey of Youth (NLSY).

greater increase in mean income between ages 25 and 55—from $41 thousand to $73 thousand—but this difference is still well below what we observe in IRS data. In stark contrast, the differences in paid-employed mean incomes across IRS and both CPS samples shown in Panel D of Figure 1 are small across ages.[20]

The CPS and IRS differences in mean incomes for the self-employed would be even more striking if we were to make adjustments for misreporting on income tax returns. Using audit data from the IRS, the BEA adjusts net profits of nonfarm proprietorships, partnerships, and S corporations when compiling the U.S. national accounts. The implied net misreporting rate over our sample period is 46 percent of the estimated true income in the case of proprietors and partners and 14 percent of true income in the case of S corporation owners. If owners were to misreport similarly on both tax forms and the CPS surveys, the magnitudes in Figure 1 would differ, but the gaps would remain the same. However, findings of Imboden, Voorheis, and Weber (2023) suggest that the true mean gaps are larger. In a matched CPS-IRS sample of self-employed individuals running unincorporated businesses, they find that on average the CPS respondents reported 50 percent *more* on the survey than on tax filings.

Finding 3. IRS data show that the main differences between IRS and CPS self-employed income distributions stem from the top and bottom deciles.

To investigate the discrepancies in the left panels of Figure 1, we compute distributional moments and owner characteristics for the two CPS samples—with and without wages of incorporated owners reclassified—and compare them to our three IRS samples from Table 1. In Table 3, we report distributional statistics for total and self-employment incomes as well as select characteristics.[21] The first three columns show summary statistics for the self-employed samples underlying Figure 1. From the counts, we see that categorizing incorporated owners as self-employed increases the sample significantly from 85 million to 138 million, but both sample counts are well below 169 million for the IRS sample. Aggregating across ages, we again find that median total incomes are lower in the IRS than in the CPS, and mean total incomes are higher in the IRS than in the CPS. The same is true for the IRS balanced panel samples shown in the last two columns. If we compare incomes at the 90^{th} percentiles, the differences are stark: the right tails for the CPS sample are significantly thinner than all three IRS samples.

The fact that the self-employment income distribution is right skewed means that the typical dollar in self-employment does not come from the typical self-employed individual. To see this more directly, consider using the income cutoffs for the first CPS sample in Table 3 to compute shares by count and income for all CPS and IRS samples. The results of this exercise are reported in Table 4. In the first CPS sample, the top 10 percent by count earns above $100 thousand. In terms of income, these individuals account for 42 percent. With incorporated owners included, individuals earning above this threshold are 16 percent of the sample and account for 53 percent of the income. Individuals in the IRS sample above the $100 thousand threshold account for roughly 80 percent of income, regardless of which sample we use.

To investigate the main sources of the IRS-CPS mismatch, we compute the fraction of the difference in mean incomes that is attributed to the right tail of the distribution. We proceed as follows. Define a threshold income level y_g^{th} so that it represents the top percentile of the CPS distribution of income for self-employed for a particular industry $g \in \mathcal{G}$. Let $\{w_g^{\text{S}}, \bar{y}_g^{\text{S}}\}_{S \in \text{IRS,CPS}}$ be

[20] Bollinger et al. (2019) compare average CPS and SSA W-2 earnings for 440 thousand individuals that appear in both samples. Over the period 2005–2010, they find the difference in average income is $813 (reported in 2010 dollars).

[21] For the paid-employed, we report the same statistics for the CPS and IRS comparison shown here and below in the online appendix.

Table 3: CPS and IRS Self-Employed Sample Comparison

	CPS Samples		IRS Samples		
Statistic	(1)	(2)	(1)	(2)	(3)
Observations (Mil.)	85.1	138.3	169.2	97.7	107.6
Incomes (2012$, Th.)					
Mean, total income	49.6	66.2	74.1	85.4	88.6
Percentiles, 10^{th}	8.5	10.6	0.0	5.5	5.9
25^{th}	15.7	20.4	10.5	11.8	12.7
50^{th}	30.1	38.9	22.0	26.4	29.0
75^{th}	54.6	74.0	60.7	73.9	78.4
90^{th}	100.4	138.1	159.9	190.3	193.3
Mean, SE income	47.6	65.2	70.3	80.9	75.0
Percentiles, 10^{th}	8.4	10.4	−5.0	5.4	−6.0
25^{th}	15.3	20.2	10.0	11.0	10.2
50^{th}	29.6	38.8	21.0	25.1	23.1
75^{th}	53.2	72.9	57.7	69.8	64.1
90^{th}	98.2	135.8	152.0	180.8	167.8
College educated (%)	40.7	46.8	NA	54.8	56.5
Top NAICS codes					
1^{st}	23	23	23	23	23
2^{nd}	81	44	54	54	54
3^{rd}	44	54	81	81	81
4^{th}	56	81	62	44	44
5^{th}	62	62	44	62	62
Demographics					
Male (%)	64.9	68.0	72.6	72.7	73.0
Married (%)	70.4	74.4	62.5	66.0	67.2
Birth year	1961	1961	1961	1962	1962

Notes: SE=self-employed and PE=paid-employed. Summary statistics are reported for two CPS samples that differ in the criteria used to assign individuals in a particular year as "self-employed." CPS sample (1) uses business income reported in the survey and categorizes an individual as self-employed if the absolute value of the income is greater than $5,000 (in 2012 dollars) and greater than income from paid-employment. The same criteria are used for the self-employed for CPS sample (2) *after* reclassifying the wage and salary income of incorporated business owners as income from self-employment. The criteria used to assign individual-year observations in the IRS self-employed sample (1) are the same as for the CPS sample (1). Sample criteria for IRS samples (2) and (3) are the same as in Table 1. To ensure that no confidential information is disclosed, reported IRS percentiles are computed as an average of observations around the value listed in the table.

Table 4: CPS and IRS Self-Employed Shares (%)

CPS Percentiles	Income Cutoff	CPS Samples (1)	CPS Samples (2)	IRS Samples (1)	IRS Samples (2)	IRS Samples (3)
		Shares, by count				
5^{th}	5,800	4.9	3.2	11.8	10.5	9.8
10^{th}	8,500	5.1	3.7	6.5	5.4	5.0
25^{th}	15,700	15.0	11.1	20.9	18.6	17.4
50^{th}	30,000	25.0	20.8	19.4	19.0	18.7
75^{th}	54,600	24.9	25.4	14.2	14.9	15.6
90^{th}	100,100	15.1	19.8	11.1	12.3	13.5
95^{th}	158,500	5.0	8.0	6.0	7.0	7.5
Above	–	5.0	8.0	10.1	12.2	12.5
		Shares, by income				
5^{th}	5,800	−0.5	−0.2	−8.8	−6.8	−5.9
10^{th}	8,500	0.7	0.4	0.6	0.5	0.4
25^{th}	15,700	3.7	2.0	3.4	2.6	2.4
50^{th}	30,000	11.5	7.2	5.7	4.9	4.6
75^{th}	54,600	20.5	15.8	7.8	7.1	7.2
90^{th}	100,100	22.1	21.9	11.1	10.7	11.3
95^{th}	158,500	12.4	15.0	10.2	10.3	10.6
Above	–	29.7	37.9	70.0	70.7	69.4

Notes: SE=self-employed and PE=paid-employed. See Table 3 for details on the self-employed CPS and IRS samples. The table reports shares—in individual counts and total incomes—after ranking individuals by their total incomes. Income bins are defined using percentiles of income for the CPS sample (1). The first row of the top panel is the share of self-employed individuals in the different samples with total income less than 5,800 in 2012 dollars, and the first row of the bottom panel is the total income share for these individuals. The bottom rows in each panel are shares by count and income for those with total income above 158,500.

the weights and means of self-employed individual-year pairs whose income exceeds y_g^{th} for industry g in the two samples. The fraction of the difference in mean incomes that is attributed to the right tail of the distribution is then given by

$$\frac{\sum_{g\in\mathcal{G}}\left(w_g^{\text{IRS}}-w_g^{\text{CPS}}\right)\frac{\bar{y}_g^{\text{IRS}}+\bar{y}_g^{\text{CPS}}}{2}+\sum_{g\in\mathcal{G}}\left(\bar{y}_g^{\text{IRS}}-\bar{y}_g^{\text{CPS}}\right)\frac{w_g^{\text{IRS}}+w_g^{\text{CPS}}}{2}}{\bar{y}^{\text{IRS}}-\bar{y}^{\text{CPS}}}. \tag{2}$$

The first term in the numerator is the total difference in means attributed to group representation—or weights ω_g—and the second term is the total difference attributed to group income means $\bar{y}_g$. The denominator is the total difference in means.

Regardless of which IRS and CPS samples are compared, we find that the fraction of the difference in mean incomes due to mismatches in the right tail of the income distribution is the dominant term.[22] In fact, in all sample comparisons, we find that the ratio in equation (2) is well

[22] Abraham et al. (2020) analyze a datafile that links IRS data for individuals with *any* self-employment over the

over 100 percent because the average income of top earners is much higher in the IRS samples than in the CPS samples. To illustrate this, consider comparing our main IRS sample—IRS sample (3)—with the CPS sample that includes incorporated owners with the self-employed (that is, IRS sample (3) versus CPS sample (2) in Table 4). If we use the income threshold of $100 thousand as before, we find the average income of those above that is $355 thousand in the IRS sample, which is much higher than the average of $218 thousand in the CPS sample. Furthermore, we find a higher mean income in the right tail for every industry. If we compute the ratio of the sum of differences in mean income—the second term in the numerator of equation (2)—relative to the sum of differences in mean income plus weights—we find an estimate of 69 percent. We should note, however, that while the difference in overall mean income is due in large part to the overall difference in mean income in the right tail, there are issues with group representation in many of the industries, but $\omega_g^{IRS} - \omega_g^{CPS}$ is sometimes positive and sometimes negative. In terms of industries, professional services and health care contribute the most to the IRS-CPS mismatch.[23] This is true regardless of the percentile we use for the income threshold and the combination of IRS and CPS samples we compare.

Next, consider the left tail of the IRS and CPS distributions. If we compare the CPS samples with the comparable IRS sample (1), which were used to create Figure 1, we find much larger shares below the income cutoff of $6 thousand for the IRS, regardless of whether we use counts or income shares. By count, the IRS sample has a share of 12 percent of self-employed individuals, whereas the two CPS samples have 5 percent and 3 percent, respectively. By income, the IRS sample has a share of -9 percent of total income for the self-employed, whereas the two CPS samples are close to zero.

An important source of the differences in statistics for the left tail are owners with net losses. If we count person-year observations with losses in the CPS, we find the fraction of self-employed with losses is 2 percent and 3 percent, respectively, for the two CPS samples in Figure 1. Conditional on having a loss, the mean loss is $10 thousand. If we compute the same statistics for the IRS population sample shown in Figure 1, we find the fraction of self-employed with losses is 2 percent, and conditional on having a loss, the mean loss is $66 thousand.

The cross-sectional statistics we have explored thus far are useful for highlighting potential issues with survey data. However, to investigate the nature of entrepreneurship—the earnings profiles over the life cycle and the determinants of entrepreneurial choice—we turn next to our study of the longitudinal panel of IRS administrative data.

4 Life-cycle Earnings Profiles

In this section, we describe and motivate the statistical model and estimation procedure that we use to estimate growth in income over the life cycle. Our method exploits the presence of multiple cohorts to separately estimate age and time effects for disaggregated subgroups within employment status. For now, we describe the procedure for an arbitrary assignment of individuals to groups and later describe how we construct the groups.

$600 threshold for filing Schedule SE to their CPS records and find higher average self-employment income in the IRS when compared to the CPS. Two key differences in our sampling lead to the opposite conclusion: namely, we use more restrictive income thresholds when categorizing individuals as self-employed, and we include S corporation owners.

[23] In his analysis of data from the Survey of Income and Program Participation, Hamilton (2000) excluded lawyers and doctors because of reclassification and top-coding concerns but noted that there were few in the SIPP dataset.

4.1 Econometric Framework

Let $i \in \mathcal{I}$ be a set of individuals; $t \in \mathcal{T} = \{t_0, t_0+1, \dots t_0+T\}$ be a set of calendar dates; $c \in \mathcal{C} = \{c_0, c_0+1, \dots c_0+C\}$ be a set of birth years (or *cohorts*); $a \in \mathcal{A} = \{a_0, a_0+1, \dots a_0+A\}$ be a set of ages; and $g \in \mathcal{G}$ be a set of observable time-invariant characteristics (or *groups*) that partition $\mathcal{I}$. Let $y_{i,t}$ be the income of individual i at date t. We use $a(i,t)$ to denote the age of individual i at date t, $g(i)$ to denote the group of individual i, and $c(i)$ to denote the cohort of individual i.

We define two functions $\beta : \mathcal{G} \times \mathcal{T} \to \mathcal{R}$ and $\gamma : \mathcal{A} \times \mathcal{G} \times \mathcal{C} \to \mathcal{R}$, which capture time, age, and cohort effects. We use the notation $\beta_{g,t}$ and $\gamma^a_{c,g}$ to denote the values of these functions for a particular collection of $\{g,t,a,c\}$, and $\beta_{g(i)}$ and $\gamma^{a(i,t)}_{c(i),g(i)}$ to be the values associated with an individual-time pair (i,t). Consider the following specification for income:

$$y_{i,t} = \alpha_i + \beta_{g(i),t} + \sum_{a=a_0}^{a=a(i,t)} \gamma^a_{c(i),g(i)} + \epsilon_{i,t}, \tag{3}$$

where $\epsilon_{i,t}$ is a disturbance term for individual i at date t. The model for income in equation (3) is quite rich. It has three components. First, the parameters $\{\alpha_i\}$ are the unobservable individual-level fixed effects that capture permanent aspects of latent ability, family inputs, and preferences as well as level effects tied to birth cohorts. We impose no restrictions on how these characteristics are distributed in the population or correlated with observable groups. Second, the parameters $\{\beta_{g,t}\}$ are the time effects that vary by calendar time and differ across groups. These parameters capture effects on income such as business cycle fluctuations. Third, the parameters $\left\{\gamma^a_{c,g}\right\}$ are the age effects that vary by age, cohort, and group. We are particularly interested in variations across subgroups based on employment status and other characteristics such as skills, industry, and demographics.

It is well-known and easy to see that one cannot separately identify β and γ. For instance, for a fixed group g, adding a constant to all $\gamma^a_{c,g}$ for which $c + a = t$ is observably indistinguishable from adding the same constant to all $\beta_{g,t}$. To make progress, we impose the following condition.

Condition 1. Age effects are the same across cohort bins of size $N_c \geq 2$.

Below, we use the notation $\overline{\gamma}^a_g$ to indicate the age effect of a group g, which is now modified to include a specification of the cohort bin, say, individuals born in the 1950s, 1960s, or 1970s. It is worth pointing out that while we impose the restriction that the age effects for sets of cohorts are the same, we impose no restrictions on how cohorts affect the level of income. The differences in mean income by cohort are absorbed in the fixed effect for individual i, namely, α_i. Condition 1 allows us to exploit the overlapping structure of our data to separate age effects from time effects.

Next, we derive the formulas needed to implement the estimation procedure. Let Δ be the time difference operator so that $\Delta x_t = x_t - x_{t-1}$. Apply Δ to equation (3) to obtain

$$\Delta y_{i,t} = \Delta \beta_{g(i),t} + \overline{\gamma}^{a(i,t)}_{g(i)} + \Delta \epsilon_{i,t}. \tag{4}$$

We work with differences in levels rather than in logarithms, given that many businesses incur losses and owners' income y_{it} can be negative.[24] To estimate the age and time effects, we propose

[24] Some authors, such as DeBacker, Panousi, and Ramnath (2022), estimate error components models after applying a transformation to income mapping $y \to \ln(y + \sqrt{1+y^2})$ to accommodate negative values of y. We view such transformations unsuitable for our purposes. To see why, consider two transitions for business owners: one from a

the following least squares problem:

$$\min_{\{\Delta\beta_{g,t},\overline{\gamma}_g^a\}} \sum_{g\in\mathcal{G}}\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} \left(\Delta y_{i,t} - \Delta\beta_{g(i),t} - \overline{\gamma}_{g(i)}^{a(i,t)}\right)^2. \tag{5}$$

By examining the first-order conditions of this minimization problem, we can better understand how the estimator works. Let $N_{g,t}^a$ be the number of individuals of group g, age a, at calendar date t. Let

$$\begin{aligned}
\overline{\Delta y}_{g,t} &= \frac{\sum_{i\in\mathcal{I}:g(i)=g}\Delta y_{i,t}}{\sum_{a\in\mathcal{A}} N_{g,t}^a} \\
\overline{\Delta y}_g^a &= \frac{\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}:a(i,t)=a,g(i)=g}\Delta y_{i,t}}{\sum_{t\in\mathcal{T}} N_{g,t}^a}
\end{aligned}$$

be the average income growth for group g between dates $t-1$ and t and the income growth averaged across time for individuals in group g between ages $a-1$ and a, respectively. We can rearrange the optimality conditions to get

$$\overline{\gamma}_g^a = \overline{\Delta y}_g^a - \sum_{t\in\mathcal{T}}\left(\frac{N_{g,t}^a}{\sum_{j\in\mathcal{T}} N_{g,j}^a}\right)\underbrace{\left\{\overline{\Delta y}_{g,t} - \sum_{k\in\mathcal{A}}\left(\frac{N_{g,t}^k}{\sum_{\ell\in\mathcal{A}} N_{g,t}^\ell}\right)\overline{\gamma}_g^k\right\}}_{\Delta\beta_{g,t}}. \tag{6}$$

Equation (6) expresses $\{\overline{\gamma}_g^a\}$ as linear combinations of two summary statistics of data, $\{\overline{\Delta y}_g^a\}$ and $\{\overline{\Delta y}_{g,t}\}$ with weights $\{N_{g,t}^a\}$. Specifically, the age effects for some age a are given by the average income growth $\overline{\Delta y}_g^a$ for that age minus an appropriately weighted average of the time effects $\{\Delta\beta_{g,t}\}$. The weights that appear in the adjustment correct for the possibility that the age distribution could be changing over time, which is relevant in our sample period.

To understand the intuition for the adjustment term in (6), consider the case in which the age distribution is constant across time, that is,

$$\frac{N_{g,t}^a}{\sum_{a\in\mathcal{A}} N_{g,t}^a} = \frac{\overline{N}_g^a}{\sum_{a\in\mathcal{A}}\overline{N}_g^a}, \tag{7}$$

where $\overline{N}_g^a = \sum_{t\in\mathcal{T}} N_{g,t}^a$. With some algebra, we can show that $\overline{\gamma}_g^a = \overline{\Delta y}_g^a - \overline{\Delta\beta}_g$, where $\overline{\Delta\beta}_g = \sum_{t\in\mathcal{T}}\Delta\beta_{g,t}/T$ is the average of time effects for group g. It simply says that the estimate of the age effect equals the average income growth for that age minus a simple average of the time effects. However, equation (7) does not hold in typical panel datasets, and therefore the second term on the right-hand side of equation (6) gives the appropriate adjustment.[25]

We make two more observations about equation (6). First, the age effect $\overline{\gamma}_g^a$ can be estimated separately for each group g. Second, one can show that the rank of the system formed by stacking equation (6) for each age is $A-1$. Therefore, we need an additional restriction—one for each group—to solve for the age effects $\{\overline{\gamma}_g^a\}$ uniquely. Following Hall (1968) and Deaton (1997), we impose the following condition.

loss of \$1,000 to a gain of \$1,000, and another from a gain of \$1,000 to a gain of \$3,000. This transformation would imply that the first transition is 14 times larger than the second. Transitions from losses to gains are sufficiently common in our sample that our mean income-age profiles would be distorted by such a transformation and we prefer to work with differences in levels, which we find more interpretable.

[25] In our sample, we have a balanced panel, and therefore the mean age is necessarily increasing in calendar time as the population is aging.

Condition 2. The average time effect satisfies

$$\frac{\overline{\Delta\beta}_g}{\overline{y}_{g,t_0}} = \frac{\mu_g}{T}\sum_t (1+\mu_g)^t \tag{8}$$

for some pre-determined constant μ_g, where $\overline{y}_{g,t} = \sum_{i\in\mathcal{I}:g(i)=g} y_{i,t} / \sum_{a\in\mathcal{A}} N^a_{g,t}$ is the average income for group g in year t and t_0 is the first year of the sample.

Condition 2 allows the estimation to match the cyclical variation in the time effect across groups in a flexible way. This is especially helpful in our sample given the severe economic downturn in 2008–2009. In particular, we do not need to take a stand on the differential effects of aggregate shocks on groups.

4.2 Groups

To implement the approach sketched out in the previous section, we need to define groups. A *group* is a Cartesian product of time-invariant characteristics that we call *subgroups.* In our case, there are 32,256 subgroups. Some of the characteristics are inherently time invariant, such as gender and birth cohort, and others are time varying such as marital status. In this section, we provide a summary of the subgrouping.

We start with education, skill, and industry. Our classifiers described in Section 2.3 are used to assign probabilities that an individual has a college education and certain skills. Given that we start at age 25, we treat these likelihoods as invariant traits. The subgroup *college educated* has two values: 1 if the education classifier is above the 0.5 cutoff and 0 if not. Similarly, the subgroups *cognitive*, *interpersonal*, and *manual* each take on one of two values: 1 if the skill value is above the 0.5 cutoff and 0 if not. The subgroup *industry* is the two-digit NAICS code for an individual's primary industry and takes on 21 possible values (including "missing"). For the time-invariant industry assignment, we use the code observed in most sample years.

For demographic characteristics, we have two inherently time-invariant traits, gender and birth cohort, and two that vary, married and children. The subgroup *gender* has two values for male and female. The subgroup *cohort* has three values: "1950s" if born between 1950 and 1959, "1960s" if born between 1960 and 1969, and "1970s" if born between 1970 and 1975. Since we are working with a balanced panel, we observe a significant overlap of cohorts over time, namely, 26 birth years (1950–1975), across 41 ages (25–65) and 16 calendar years (2000–2015). The subgroup *married* has two values: 1 if the individual is married for nine or more years in the sample—not necessarily to the same person—and 0 otherwise. The subgroup *children* has two values: 1 if the individual claims children on their tax return at any time in the sample and 0 otherwise.

Finally, we include a time-invariant subgrouping related to employment status. In Section 2.2, we assigned an employment status of self-, paid-, or non-employed to each individual-year observation. Here, we study individuals over their life cycle with a particular focus on individuals that are engaged in market work in most years of our sample. We assign individuals to a subgroup *primarily employed* if they have 12 or more years in either self- or paid-employment—with any number of switches between these two—and at most one intermediate year of non-employment between years of either self- or paid-employment. Other non-employment years could occur at the beginning or end of the sample period. For example, the primarily employed subgroup would include individuals that were still in college at the beginning of the sample but were subsequently employed and individuals that retired early but were previously employed—as long as the number of years in some form of employment over the sample is at least 12. All other individuals are assigned to the complementary subgroup that we call *not primarily-employed.*

To address key questions about the nature of entrepreneurship, we disaggregate the primarily employed group further in order to analyze certain subgroups of interest. We first split the primarily employed group into two mutually exclusive subgroups: individuals that are *primarily paid-employed* and individuals that have *tried self-employment*. The primarily paid-employed are those with at least 12 or more years of paid-employment. These paid-employed individuals serve as a benchmark for our analysis of entrepreneurs in the tried-self-employment group. The years in self-employment need not be consecutive, thus allowing for individuals that switch in and out of self-employment one or more times in the sample period. With these switchers in mind, we further split the tried-self-employment group into two mutually exclusive subgroups: the *primarily self-employed* that have 12 or more years in self-employment—like their primarily paid-employed peers—and the remaining group of *mostly switchers*, who have fewer than 12 years in self-employment.

In sum, we have a Cartesian product of 32,256 time-invariant subgroups: 16 categorized by education and skill; 21 by industry; 24 by demographics; and 4 by employment status.

4.3 Parameter Estimates

In this section, we present selected results from the estimation procedure for individuals in the tried-self-employment and primarily paid-employed groups (and refer readers to the online appendix for full results). Our econometric approach delivers estimates for age and time effects, $\{\bar{\gamma}_g^a, \Delta\beta_{g,t}\}$, for all time-invariant groups g, by age a and year t. Here, we aggregate those estimates for the two main groups of individuals that constitute all primarily employed over our sample: those in the tried self-employment subgroup and those in the primarily paid-employed subgroup.

In Tables 5 and 6, we report the aggregated parameter estimates for age and time effects, respectively. The self-employed subsample and the reference group of paid-employed have two noteworty differences. First, the estimated age effects are significantly higher for the self-employed as compared to their paid-employed peers between ages 25 and 50. For example, at age 26, the estimated age effect, $\bar{\gamma}$, is 63 percent higher for the self-employed compared to the paid-employed and remains higher until age 55. Second, the estimated time effects for the self-employed are much more volatile—and significantly lower during the Great Recession which occurs midway through the sample.

We use the estimates of the age and time effects to construct economically interpretable integrated income profiles for different aggregated groups of interest. Let $\mathcal{G}$ be one of these aggregated groups. To compute the integrated income profile for $\mathcal{G}$, we start by computing the average income of 25-year-olds in each subgroup $g \in \mathcal{G}$—call this $Y_g(25)$—and the weighted average income at age 25: $Y_{\mathcal{G}}(25) = \sum_{g\in\mathcal{G}} N_g^{25} Y_g(25) / \sum_g N_g^{25}$. We then integrate estimated growth rates and add them to the average income for 25-year olds:

$$Y_{\mathcal{G}}(a) = Y_{\mathcal{G}}(25) + \sum_{j=26}^{a} \sum_{g\in\mathcal{G}} \frac{N_g^j}{\sum_k N_k^j} (\bar{\gamma}_g^j + \overline{\Delta\beta_g}). \tag{9}$$

Finding 4. Estimates of age and time effects based on the IRS balanced panel show that individuals categorized as having tried self-employment have significantly higher average income and steeper income growth profiles than paid-employed peers with similar characteristics.

Results for those that tried self-employment and those that are primarily paid-employed are shown in Figure 2. At age 25, the profiles are similar but are much steeper over the life cycle for the self-employed. By age 55, our estimate is an average income of $134 thousand in 2012 dollars—much higher than the estimate of $79 thousand for the paid employed. These differences in profiles

Table 5: Estimates of Age Effects (2012$)

Age	Self-Employed	Paid-Employed
26	5,242	3,211
27	5,405	2,957
28	5,189	2,712
29	5,000	2,498
30	4,990	2,286
31	4,552	2,185
32	4,910	2,081
33	5,142	1,858
34	5,060	1,858
35	4,603	1,689
36	4,218	1,536
37	3,939	1,412
38	3,750	1,281
39	3,743	1,228
40	3,538	1,171
41	3,220	1,024
42	2,542	921
43	2,862	735
44	1,417	602
45	1,744	461
46	744	292
47	894	194
48	685	70
49	775	−17
50	26	−316
51	148	−426
52	−666	−532
53	−1,000	−563
54	−1,165	−833
55	−1,432	−692
56	−2,042	−1,202
57	−3,443	−1,377
58	−2,786	−1,482
59	−3,517	−1,737
60	−4,632	−2,361
61	−4,154	−3,169
62	−6,410	−4,239
63	−6,510	−5,786
64	−7,449	−4,871
65	−8,145	−5,464

Notes: Age effects ($\bar{\gamma}_g^a$) are reported in 2012 dollars for two aggregated groups: individuals in the tried-self-employment subgroup (or "self-employed") and individuals in the primarily paid-employed subgroup (or "paid-employed"). See details of the group assignment in Section 4.2 and the estimation procedure in Section 4.1.

Table 6: Estimates of Time Effects (2012$)

Year	Self-Employed	Paid-Employed
2001	1,353	721
2002	−686	255
2003	7,341	895
2004	−775	1,322
2005	5,416	1,055
2006	3,857	1,575
2007	−1,528	1,758
2008	−9,655	−373
2009	−8,785	−1,583
2010	1,661	883
2011	2,245	667
2012	10,306	672
2013	−2,846	−188
2014	4,579	1,123
2015	3,939	1,416

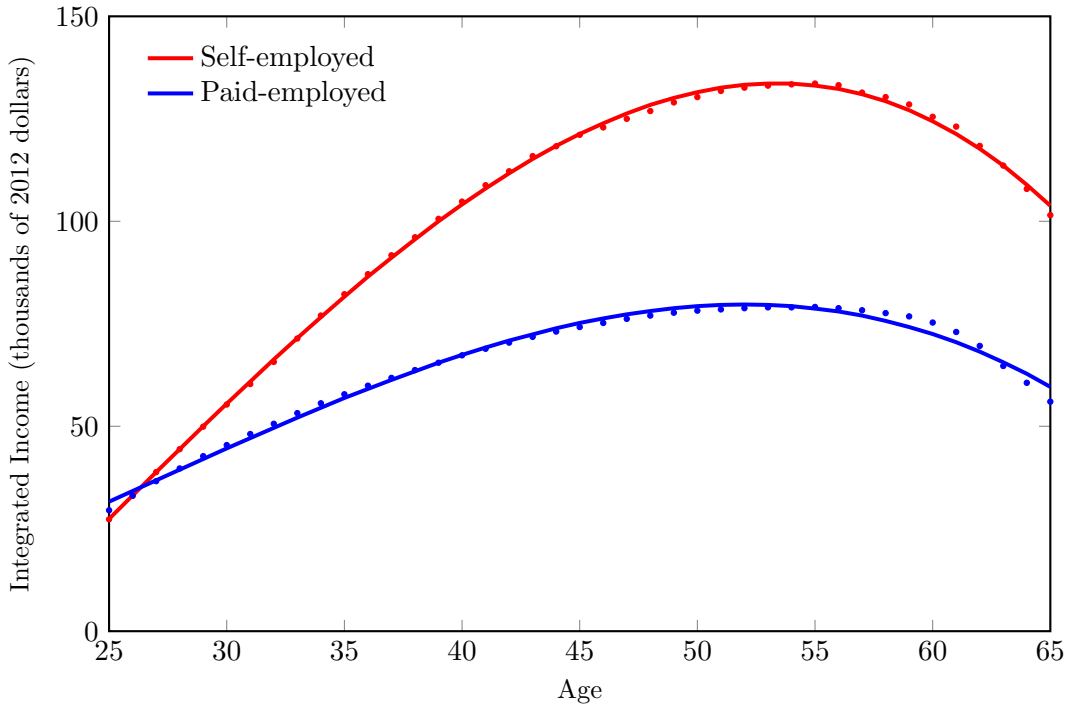
Notes: Time effects ($\Delta\beta_{g,t}$) are reported in 2012 dollars for two aggregated groups: individuals in the tried-self-employment subgroup (or "self-employed") and individuals in the primarily paid-employed subgroup (or "paid-employed"). See details of the group assignment in Section 4.2 and the estimation procedure in Section 4.1.

for the self- and paid-employed would be even more striking if we were to adjust reported incomes to account for business income underreporting. As we noted earlier, the BEA adjusts total business incomes in the U.S. national accounts using net misreporting rates of 46 percent for unincorporated owners and 14 percent for S corporation owners.

In the online appendix, we include results similar to those in Figure 2 that break out the main subgroups of the tried self-employed and include the integrated incomes of the not primarily employed. The results show that *both* self-employed subgroups have steeper profiles than their paid-employed peers. On the other hand, the not primarily employed group has a much flatter profile than any other group. We also report results in the online appendix establishing that steeper income growth profiles of the self-employed when compared to the paid-employed is a general pattern. For example, we can compare men that tried self employment to those with little to no self-employment experience. The self-employed men have significantly steeper income profiles than either the paid- or non-employed groups. The estimates at age 55 are $187 thousand for those primarily self-employed with 12 or more years of self employment and $123 thousand for those switching between self- and paid-employment. On the other hand, the estimates for the reference groups at age 55 are $92 thousand in the case of the primarily paid-employed and $19 thousand in the case of the not primarily employed. While the exact estimates vary for other characteristics, this pattern repeats itself in all subgroups but one: those categorized as not likely to be college educated. In that case, the self-employed profiles are not higher but are relatively close.

While estimates of the age effects are most critical for constructing the life-cycle income profiles, estimates of the time effects provide a useful gauge of the aggregate risks in self-employment,

Figure 2: Estimated Income Profiles for Self- and Paid-Employed



Notes: The figure shows integrated income $Y_{\mathcal{G}}(a)$ at each age a for two aggregated groups $\mathcal{G}$: individuals in the tried-self-employment subgroup (or "self-employed") and individuals in the primarily paid-employed subgroup (or "paid-employed"). Integrated income is defined in equation (9). See details of the group assignment in Section 4.2 and the estimation procedure in Section 4.1.

especially since the Great Recession occurs during our sample period.

Finding 5. Estimates of time effects based on the IRS balanced panel show that individuals categorized as having tried self-employment had steep declines in growth during the Great Recession compared to paid-employed peers with similar characteristics.

In Figure 3, we plot the time effects relative to average income for individuals in the two main groups, that is, a weighted sum of $\Delta\beta_{g,t}$ divided by $\overline{y}_{g,t}$ for each each subgroup $g \in \mathcal{G}$, where $\mathcal{G}$ is either tried self-employment or primarily paid-employment and weights are constructed by group counts. The figure shows large differences in growth for the two groups during the 2008–2009 downturn. Incomes were down -10 percent of average income for self-employed individuals in 2008 and imperceptibly for the paid-employed. In 2009, the self-employed experienced further declines, with incomes falling -9 percent, while the paid-employed experienced a relatively modest decline.

4.4 Non-pecuniary Benefits of Entrepreneurship

The findings above are ostensibly inconsistent with the common portrayal of the self-employed individual choosing entrepreneurship for non-pecuniary reasons. According to Hurst and Pugsley (2011), over 50 percent of new business owners cite benefits like flexible schedules and being one's

Table 7: Statistics After Grouping Individuals in IRS Balanced Panel

Statistic	Primarily Self-Employed			Mostly Switching SE/PE	Primarily Paid-Employed	Not Primarily Employed
	Total	Earning Above PE	Earning Below PE			
Individuals (Mil.)	2.8	1.2	1.6	3.4	47.2	26.6
Shares (%)						
Counts	3.4	1.5	1.9	4.3	59.0	33.3
Total income	9.4	7.9	1.5	7.3	72.2	11.1
SE income	62.3	52.6	9.5	21.4	5.3	11.0
PE income	1.4	1.1	0.3	5.1	82.3	11.1
Incomes (2012$, Th.)						
Mean, total income	134.2	262.1	37.8	83.3	60.0	16.4
Percentiles, 10^{th}	15.0	54.3	11.8	14.8	21.0	1.6
25^{th}	26.7	79.6	18.2	23.4	30.3	5.0
50^{th}	56.6	144.9	30.3	41.7	45.3	10.8
75^{th}	130.1	275.7	50.6	81.8	68.0	20.0
90^{th}	291.3	528.0	84.6	164.8	102.5	33.5
Mean, SE income	117.0	230.4	31.6	32.4	0.6	2.1
Percentiles, 10^{th}	12.3	43.1	9.6	−0.1	−0.1	0.0
25^{th}	22.1	66.4	15.2	4.5	0.0	0.0
50^{th}	47.5	124.0	25.8	11.3	0.0	0.0
75^{th}	112.3	242.8	43.8	29.7	0.0	0.6
90^{th}	256.9	469.0	74.0	73.2	1.0	4.6
Mean, PE income	17.2	31.8	6.3	50.9	59.4	14.2
Percentiles, 10^{th}	0.0	0.0	0.0	8.8	21.0	0.6
25^{th}	0.0	0.0	0.0	15.1	30.3	3.5
50^{th}	2.6	6.1	1.5	27.0	45.1	9.3
75^{th}	11.6	24.3	6.5	50.2	67.5	18.3
90^{th}	35.2	68.6	16.0	93.5	101.3	30.9
Education and skills (%)						
College educated	62.4	73.4	54.2	63.6	59.7	38.4
Cognitive	61.4	63.1	60.3	62.7	57.7	40.9
Interpersonal	60.5	71.3	52.5	63.4	65.6	45.5
Manual	36.0	29.4	40.9	36.9	36.1	40.6
Primary industry (%)						
Agriculture	1.8	2.0	1.6	1.2	0.9	1.3
Mining	0.5	0.5	0.4	0.5	0.5	0.3
Utilities	0.1	0.0	0.0	0.1	0.3	0.1

See notes at end of table.

Table 7: Statistics After Grouping Individuals in the IRS Balanced Panel (cont.)

Statistic	Primarily Self-Employed			Mostly Switching SE/PE	Primarily Paid-Employed	Not Primarily Employed
	Total	Earning Above PE	Earning Below PE			
Primary industry (%)						
Construction	17.9	15.1	20.2	13.5	5.6	8.3
Manufacturing	3.6	4.5	2.9	6.2	14.2	8.4
Wholesale trade	3.4	4.6	2.4	3.7	4.0	3.0
Retail trade	8.1	9.1	7.5	8.6	8.3	10.0
Transportation	5.1	2.8	7.0	5.8	3.1	3.3
Information	1.0	0.9	1.0	1.6	1.8	1.3
Finance	4.2	3.7	4.6	3.9	3.7	2.5
Real estate	5.0	4.6	5.4	4.1	1.8	2.7
Professional services	15.5	18.8	13.0	15.5	8.9	7.8
Management	0.1	0.1	0.1	0.2	0.9	0.4
Administration	4.3	4.1	4.5	4.6	4.0	6.6
Education	0.4	0.4	0.4	0.6	0.5	0.6
Health care	9.5	13.3	6.7	8.6	5.6	7.1
Arts	1.9	1.5	2.3	1.7	0.8	1.2
Accommodation	3.6	4.2	3.2	4.4	3.5	6.2
Other services	11.1	8.6	13.0	8.0	2.2	5.5
Other NAICS	0.7	0.3	0.6	5.0	17.1	10.1
Missing NAICS	2.3	1.0	3.3	2.1	12.2	13.3
Employees and Profits						
Ever had employees (%)	63.9	82.4	50.0	47.9	4.7	8.2
Gross profits (2012$, Th.)	400.7	732.9	147.9	156.5	6.7	19.0
Demographics						
Male (%)	81.6	79.5	83.4	72.9	52.5	39.9
Married (%)	78.0	81.8	75.4	70.3	66.8	56.1
Has children (%)	84.8	85.3	84.7	85.2	82.5	80.7
Mean number of children	2.2	2.2	2.3	2.4	2.2	2.3
Median birth year	1960	1960	1960	1964	1963	1964
Other incomes (2012$, Th.)						
Mean, spousal wages	26.3	32.1	22.0	27.3	30.6	38.0
Mean, asset income	56.0	90.8	29.6	21.4	4.6	9.5
Mean, UI income	0.1	0.1	0.1	0.4	0.4	0.5

Notes: SE=self-employed and PE=paid-employed. These summary statistics are computed using panel data for individuals in IRS sample (C) in Table 1, with the criteria for self-employed in (3). See Section 2 for more details on the sample construction and income measures. For more details on the time-invariant groups listed as the top of each column, see Section 4.2. Incomes and gross profits are reported in thousands of 2012 dollars. To ensure that no confidential information is disclosed, reported percentiles are computed as an average of observations around the value listed in the table. Industries are classified by two-digit NAICS.

Figure 3: Estimated Time Effects Relative to Average Income

Notes: The figure shows weighted averages of the estimated time effects, $\Delta\beta$, at each date for two aggregated groups: individuals in the tried-self-employment subgroup (or "self-employed") and individuals in the primarily paid-employed subgroup (or "paid-employed"). Time effect parameters for each group g in year t are divided by average income, $\overline{y}_{g,t}$. Weights used in aggregating are group counts. See details of the group assignment in Section 4.2 and the estimation procedure in Section 4.1.

own boss as primary reasons for starting their business. Hamilton (2000) estimates a median earnings differential of 35 percent in compensation for these non-pecuniary benefits. To investigate this narrative, we construct a subsample of individuals that fit the Hurst-Pugsley-Hamilton narrative: they are self-employed but earn less than paid-employed peers with the same education, skills, industries, and demographics. We compare summary statistics—counts, incomes, characteristics—for this subsample with statistics for other subgroups of individuals as we vary employment status.

Finding 6. IRS data show that many of the primarily self-employed earned less over the sample years than paid-employed peers with similar characteristics, but in the aggregate this subgroup has a much lower share of the total income than those that earned more than their peers.

In Table 7, we report the same summary statistics shown earlier in Table 2, but do so after first grouping individuals by employment status. In the first row of the table, we list counts by employment type, which aggregates to 80 million individuals (or, equivalently, 1.3 billion observations divided by 16 years). Together, individuals in the first and fourth columns constitute the group that tried self-employment. The primarily self-employed shown in the first column derive most of their income from running businesses. They are only 3.4 percent of our sample in terms of counts but earn 62 percent of self-employment income. The individuals that mostly switch between self- and paid-employment shown in the fourth column derive more income from paid-employment over

the sample but are a relatively small group with significant self-employment earnings. In the fifth column of the table, we report statistics for the primarily paid-employed individuals that serve as a reference group for those that tried self-employment.[26] The last column shows results for individuals that are not primarily employed, which is a large group in counts but small in terms of market activity.

To investigate the role of non-pecuniary benefits, we decompose the primarily self-employed group in Table 7 into two subgroups: those with an average sample income above and below their primarily paid-employed peers. More specifically, for each individual in this group, we compute the sum of their total income over the full sample and compare it to the sum of total income for primarily paid-employed peers with the same characteristics—that is, the same values for education, cognitive skill, interpersonal skill, manual skill, industry, gender, marriage, children, and birth year. We refer to the subsample of self-employed individuals with average incomes over the sample exceeding averages of their paid-employed peer group as the *earning above paid-employed* subsample. The complement group is assigned to the *below paid-employed* group. The statistics for these two groups are reported in the second and third columns of Table 7.

In terms of relative sizes of the subgroups, we find that the individuals with average incomes below paid-employed peers account for 57 percent of the 3 million individuals in the primarily self-employed group. This is the sense in which self-employed individuals—consistent with the Hurst and Pugsley (2011) and Hamilton (2000) narrative—are typical. However, they earn only 16 percent of the total income of the primarily self-employed group. This is the sense in which they are atypical: they do not earn the typical dollar in self-employment.

Because aggregate policies aimed at the self-employed affect both subgroups, we ask whether the IRS data shown in Table 7 shed any new light on potential differences between those earning above paid-employed peers and those earning below—other than measures of income or business size (for example, gross profits or employment). We find that they are surprisingly similar in terms of industrial composition and demographics and, on the whole, different from the primarily paid-employed on these dimensions. The self-employed groups are more likely to work in construction, professional services, health care, and other services, while the paid-employed work in manufacturing and retail trade. They are also demographically different: a significant fraction of individuals in the self-employed groups are married men, whereas the paid- and non-employed groups include many more women and singles. We find some differences in education and skill, with those earning below paid-employed peers less likely to be college educated or interpersonally skilled and more likely to be manually skilled. In the online appendix, we provide more detail for Table 7 by computing the statistics separately by education, gender, and an indicator of whether individuals have ever been employers. We find that less-educated males that have never been employers are the most likely to be grouped with the self-employed earning below paid-employed peers.

With better data on fringe benefits for the paid-employed and the extent of income misreporting by the self-employed, estimates for the share of owners driven by non-pecuniary motives will change, but it is unlikely that we could overturn the finding that these owners earn a small share of income for the primarily self-employed. Consider first the impact of adding in fringe benefits for the paid-employed. According to the BEA, fringe benefits are equal to 23.5 percent of employee compensation in the aggregate over our sample period. If we assume that all paid-employed earn this amount, we would have to shift some individuals from the "above PE" group in Table 7 to the "below PE" group. However, as is clear from Table 7, only a small number of individuals are on that margin. As a result, adding the paid-employed benefits to the W-2 incomes would

[26] A tiny fraction—around 25.3 percent—of the 1.3 billion observations in Table 2 are person-years in self-employment that were included with the primarily paid-employed when we constructed invariant groups.

not significantly change our estimates for the counts or income shares of those earning less than paid-employed peers.

Suppose instead that we want to account for the widespread income underreporting by business owners. Recently published data based on IRS audit compilations reported by the BEA show that 34 percent of pass-through income—roughly $700 billion dollars in 2018—was misreported. Also well documented is the fact that pass-through businesses have few reporting requirements and thus have ample opportunities for misreporting their taxable income. This is especially true for sole proprietorships. According to IRS audit estimates, roughly 55 percent of nonfarm proprietor income is not reported, and the misreporting occurs across the income distribution.[27] Suppose we re-create the subsample of self-employed individuals earning below and above their paid-employed peers after adjusting all Schedule C incomes and losses by the average IRS imputation rate.[28] In this case, we would find that the number of individuals in the below paid-employed group falls to 37 percent of the primarily self-employed group and their income share falls to 10 percent. Including misreporting by partners and S corporation owners would likely drive these estimates even lower.

We turn next to a related narrative in the literature concerned with the inherent risk of entrepreneurial activity and whether business owners are compensated for greater uncertainty in incomes.

4.5 Risk Factors in Entrepreneurship

In Section 4.3, we documented significantly steeper income profiles for self-employed individuals as compared to those who are paid-employed. This leads to a natural question of whether higher returns in self-employment are compensation for risk. The existing macro and finance literature that studies the risk-return trade-off using survey-based evidence finds a puzzling discrepancy: the measured returns for self-employed individuals do not seem to match the level of risk entrepreneurs undertake.[29] This phenomenon is known as the "private equity puzzle." To explain this discrepancy, researchers have suggested that non-pecuniary benefits discussed in Section 4.4 form a significant part of the returns to self-employment. In this section, we contribute to this discussion by examining patterns in the variability and persistence of income growth among self-employed individuals and comparing them to the patterns of paid-employed individuals. We first consider idiosyncratic risk faced by individuals and then investigate how the self-employed fared in the Great Recession of 2008–2009.

4.5.1 Idiosyncratic Risk

Addressing whether entrepreneurs' risk-taking is puzzling requires modeling consumption risk and taking a stance on insurance sources. The approach we take here is to transparently map moments in our data to statements about implied risk aversion that rationalize entrepreneurial risk-taking. We then compare our derived measures of risk aversion to estimates from the household finance literature.[30] Key moments include percentiles and rank autocorrelations of the distribution of income growth for the self- and paid-employed.

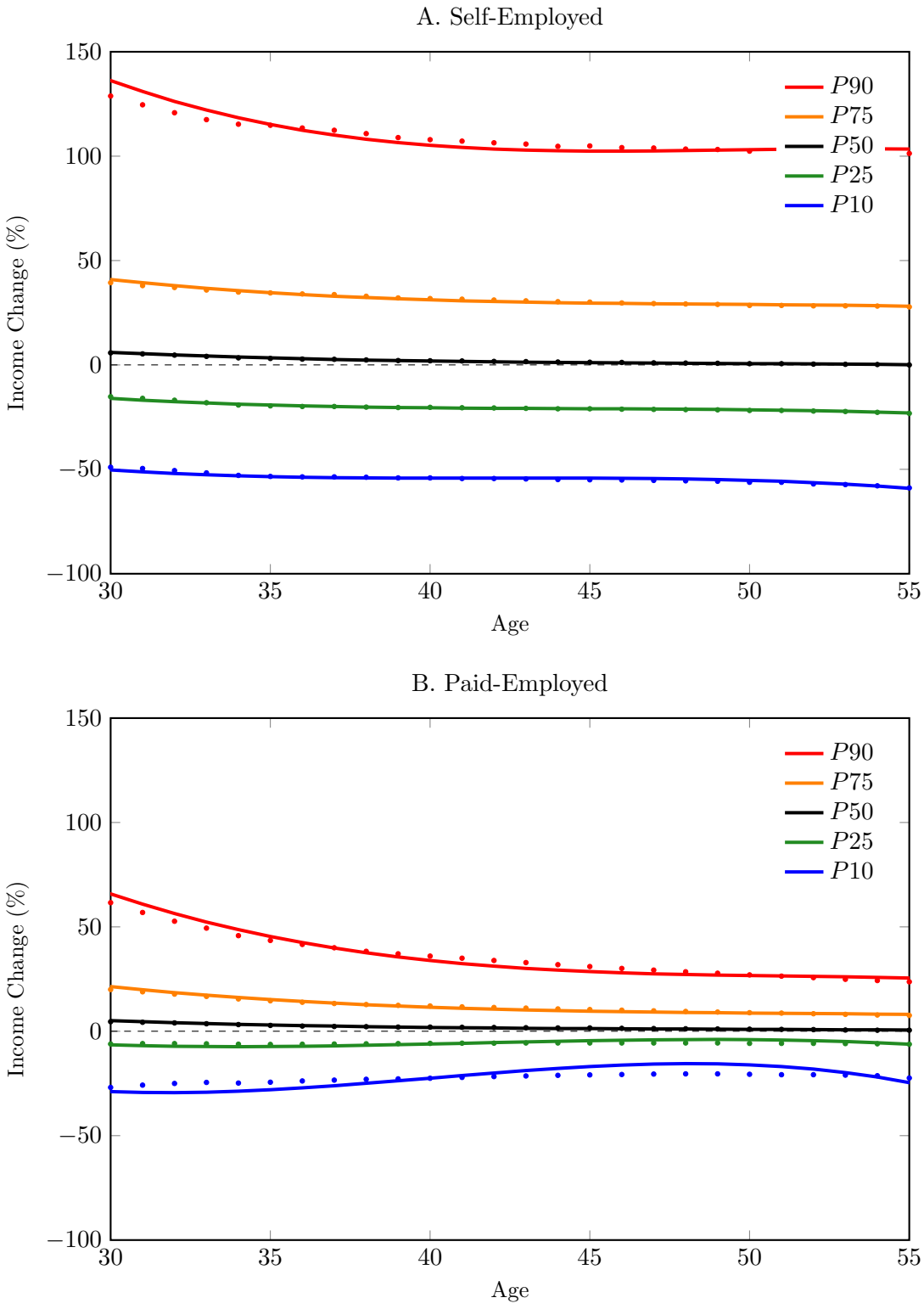
[27] See Johns and Slemrod (2010) and Auten and Langetieg (2023).

[28] If the Schedule C income is positive, we divide by 0.45 and, if it is negative, we multiply by 0.45.

[29] See Moskowitz and Vissing-Jorgensen (2002) and references therein. See also Hall and Woodward (2010), who study the universe of U.S.-based high-tech startups.

[30] These moments can also be used as inputs to a fully specified structural model of self-employed individuals from which consumption can be inferred. See, for example, Cagetti and DeNardi (2006), Bhandari and McGrattan (2021), or Catherine (2022).

Figure 4: Age-Over-Age Growth in Incomes



Notes: The figure shows selected percentiles of the distribution of age-over-age income changes by age: $\Delta y_{ia}/|y_{i,a-1}|$ for individual i at age a. To ensure that no confidential information is disclosed, reported percentiles are computed as an average of observations around the value listed in the figure. Individuals in the top panel are in the tried-self-employment subgroup (or "self-employed"), and individuals in the bottom panel are in the primarily paid-employed subgroup (or "paid-employed"). See details of the group assignment in Section 4.2.

Finding 7. IRS data show that 90-10 differences in income growth rates for self-employed are in the range of 2 to 3.5 times larger than those for paid-employed, depending on age, but the distributions for the two groups share common patterns.

In Figure 4, we plot percentiles of the distributions of income growth, $\Delta y_{i,t}/|y_{i,t-1}|$, using data for individuals indexed by i at age t pooled across all years in our sample. This estimate of growth shares properties with log differences in income but crucially allows for negative values of $y_{i,t}$, which are common for the self-employed.[31] The figure reveals several key patterns. The median ($P50$) growth rates are similar across employment groups and ages. For both the self- and paid-employed, income changes show the highest dispersion at younger ages. Self-employed incomes exhibit greater dispersion in growth rates across all age groups when compared to paid-employed incomes. For example, the $P90$–$P10$ range for self-employed individuals at age 35 is approximately 170 percent, which is about 2.5 larger than the $P90$–$P10$ range for the paid-employed. Interestingly, the $P90$–$P10$ variation remains relatively constant across middle ages for both groups. This pattern suggests that the volatility of incomes for the self-employed does not increase over the life cycle relative to incomes for the paid-employed, despite growing differences in the average levels. Finally, for both groups, we find positive Kelly skewness, defined as $(P90+P10-2\cdot P50)/(P90-P10)$, with values around 0.2 for our paid-employed subsample and notably higher values around 0.3 for our self-employed subsample.[32] These results indicate that income growth distributions are right skewed for both groups, with self-employed individuals experiencing a more pronounced rightward skewness in their income growth patterns.[33]

The patterns in Figure 4 prompt the question of whether the distributions of growth rate differentials reflect individual risk or persistent differences in growth rates across individuals. To explore this question, we conduct two sets of exercises. First, we construct percentiles of income changes for narrowly defined disaggregated groups. (These results are available in the online appendix.) For nearly all subgroups, the $P90$–$P10$ differences for the self-employed are about two to three times larger than those for the paid-employed. Second, we examine the (rank) autocorrelogram of growth rates to determine whether a significant latent persistent component exists. We find that the autocorrelations for both self-employed and paid-employed individuals decline sharply toward zero after the first lag.[34]

We next assess the relevance of these distributional moments for the private equity puzzle. Our strategy is as follows. We assume individuals' preferences regarding consumption risk are represented by Epstein and Zin (1989) preferences:

$$V_t\left(\{C_j\}_{j=t}^{\infty}\right) = \left[(1-\delta)\, C_t^{1-\rho} + \delta\left(\mathbb{E}_t V_{t+1}^{1-\psi}\right)^{\frac{1-\rho}{1-\psi}}\right]^{\frac{1}{1-\rho}}. \tag{10}$$

Given a model of consumption risk and values for (δ, ρ), we can determine what level of the risk aversion parameter ψ makes an individual indifferent between the income growth risk distributions of self- and paid-employed individuals. Our tractable model of consumption risk considers two

[31] In the online appendix, we report the income growth using a measure of growth net of age and time effects, represented as $\Delta\epsilon_{i,t}/|y_{i,t-1}|$. The main patterns are unchanged.

[32] This measure of skewness is less affected by outliers and is popular in the earnings inequality literature (see, for example, Guvenen, Ozkan, and Song (2014) or Guvenen et al. (2021)).

[33] Our finding that income growth for the paid-employed sample is positively skewed would seem to contradict the findings of Guvenen et al. (2021). While our sample of paid-employed individuals and our concept of income both differ from theirs, we attempted to construct more comparable samples and measures of income changes but always found positive Kelly skewness estimates.

[34] Abowd and Card (1989) document similar patterns for paid-employed income using the Panel Study of Income Dynamics.

sources of insurance: self-insurance through savings and external insurance from family, friends, and the government. To implement our strategy, we first partition income risk into permanent (r_{it}) and transitory components (z_{it}) by fitting the following model for income growth:

$$\frac{\Delta y_{i,t}}{|y_{i,t-1}|} = \Delta r_{i,t} + \sigma_z \Delta z_{i,t}, \quad r_{i,t} = \mu + r_{i,t-1} + \sigma_r \eta_{i,t}, \tag{11}$$

where $\{z_{i,t}, \eta_{i,t}\}$ are i.i.d. and follow a standard Gaussian distribution.[35] Under these assumptions, we can map income risk moments to the underlying parameters of the income process (11) using straightforward expressions:

$$\mu = P50,\ \sigma_z^2 = -A\left(\frac{P90 - P10}{2.56}\right)^2,\ \sigma_r^2 = \left(\frac{P90 - P10}{2.56}\right)^2 (1 + 2A), \tag{12}$$

where $P10$, $P50$, and $P90$ are the percentiles in Figure 4 and A is the rank autocorrelation of the income growth measure $\Delta y_{i,t}/|y_{i,t-1}|$. We index the income process by $\Theta^g = [\mu^g, \sigma_r^g, \sigma_z^g]$ where g is some group defined by employment status and age. For some $\underline{\Delta c} \in (-\infty, 0)$, we posit that the change in the logarithm of consumption is given by

$$\Delta c_{i,t} = \max\left\{\underline{\Delta c}, \phi_r \Delta r_{i,t} + \phi_z \sigma_z z_{i,t}\right\}, \tag{13}$$

where $c_{it} = \ln C_{it}$. Our formulation of consumption risk is motivated by Blundell, Pistaferri, and Preston (2008) and others, who show that access to private savings allows individuals to insure against transitory shocks $z_{i,t}$ better than permanent shocks $r_{i,t}$. For the residual risk, we model external sources of insurance as a lower bound of consumption growth with a threshold of $\underline{\Delta c}$.

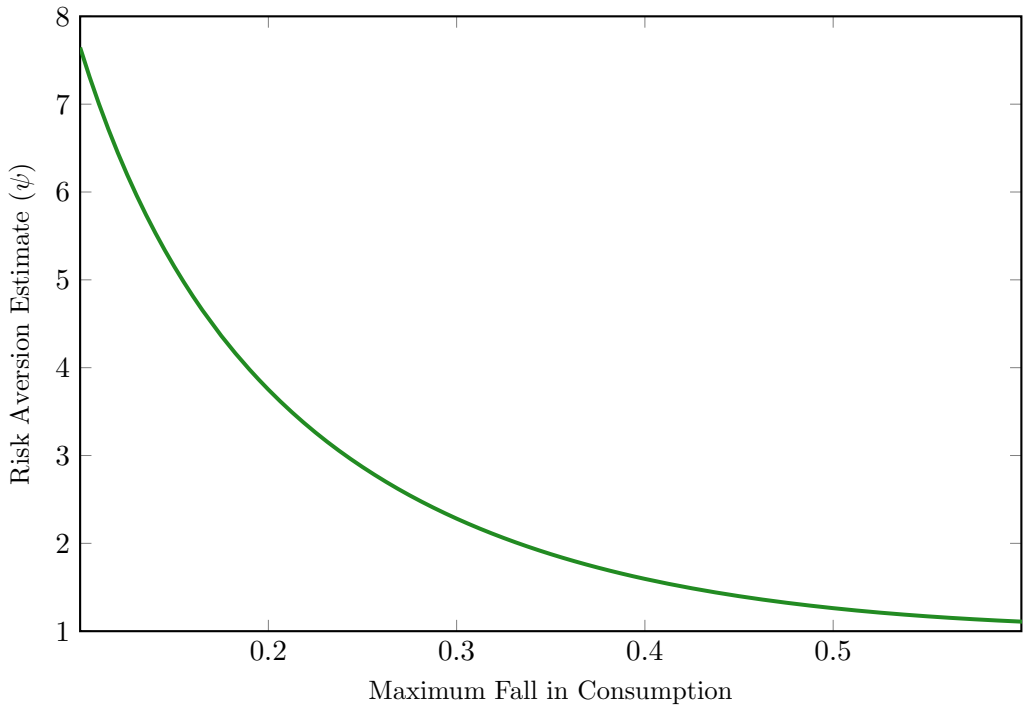
For our baseline analysis, we set $\delta = 0.96$, ρ=1, $\phi_r = 1$, and $\phi_z = 0$, with the last choice capturing the fact that transitory shocks are well insured. We report results for various values of $\underline{\Delta c}$. Figure 5 summarizes our main findings.[36] The x-axis represents $1 - \exp\{\underline{\Delta c}\}$, which indicates the maximum possible drop in consumption relative to the previous period. Higher values are associated with less insurance. The y-axis shows the required level of risk aversion ψ that makes an individual with preferences (10) indifferent between income processes $\Theta^{SE,age=35}$ and $\Theta^{PE,age=35}$. The implied risk aversion increases from around 1 to 7 as we increase the level of underlying insurance. A higher value of $\underline{\Delta c}$—to the left side of the x-axis—implies better insurance and consumption risk that is more right skewed. Keeping everything else fixed, this makes entrepreneurship a desirable prospect. In other words, to make an individual indifferent between the two employment types, the implied risk aversion needs to increase.

The literature offers a wide range of risk aversion estimates. Early studies based on household-level income and consumption data, such as Attanasio et al. (1999) and Gourinchas and Parker (2002), found estimates of around 1. More recent work in household finance that uses savings and portfolio information, such as De Nardi, French, and Jones (2010), Ameriks et al. (2020), and Calvet et al. (2021), suggests higher estimates ranging from 4 to 8. Our estimated risk aversion aligns

[35] We also allowed for the innovations to follow a skewed-Gaussian distribution and found that the main results are very similar. Our formulation of the income processes in both cases is purposefully tractable to capture key properties of income risk that are relevant for consumption risk with a small set of moments. See DeBacker, Panousi, and Ramnath (2022) for alternative models of income risk.

[36] To produce these results, we used percentile and rank autocorrelation estimates at age 35. For the self-employed, the percentile estimates for $P10$, $P50$, and $P90$—all in percentages—are −53.4 percent, 3.1 percent, and 114.8 percent, respectively, and the rank autocorrelation A is −0.18. For the paid-employed, these same estimates are −24.4 percent, 2.8 percent, 43.5 percent, and −0.13, respectively. Data for other ages and subgroups are available in the online appendix.

Figure 5: Implied Risk Aversion



Notes: The figure plots the level of risk aversion given $\delta = 0.96$, $\rho = 1$, $\phi_r = 1$, and $\phi_z = 0$ that makes an individual with preferences (10) indifferent between income processes calibrated to moments of $\Delta y_{i,t}/|y_{i,t-1}|$ for self- and paid-employment incomes at age 35. The values on the x-axis are the maximum consumption drop, that is, $1 - \exp\{\underline{\Delta c}\}$. For more details of the computation, see Section 4.5.1.

with the household finance literature for insurance levels corresponding to maximum consumption drops of 10 to 20 percent.[37] Our conclusion from this analysis is that it is possible to rationalize a more dispersed income risk distribution for self-employed individuals if one is willing to accept that individuals are insured against the most adverse shocks. Such insurance makes self-employment an attractive option and not puzzling from a risk-return perspective.[38] It also highlights the importance of measuring sources of insurance for the self-employed.

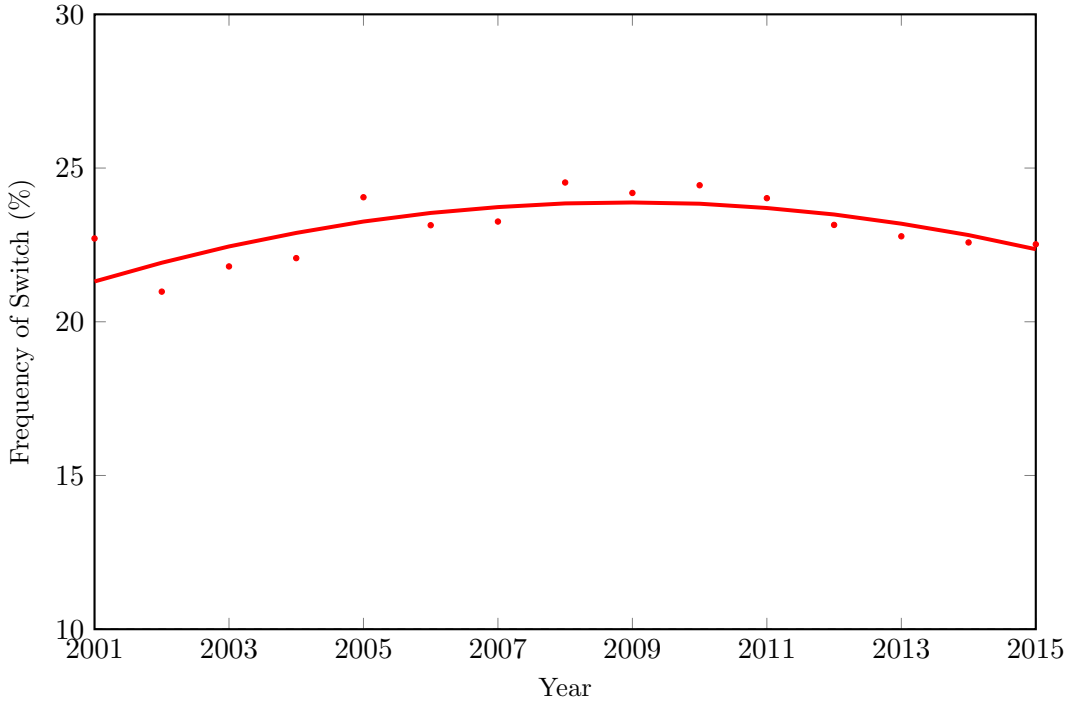
4.5.2 Aggregate Risk

In addition to idiosyncratic risks, we use our longitudinal dataset to study aggregate risk for individuals in cyclically sensitive sectors by exploiting the fact that our sample includes many years before and after the Great Recession of 2008–2009. As we showed in Figure 3, the self-employed experienced much larger declines in annual income growth as compared to the paid-employed. Here,

[37] Blundell, Pistaferri, and Preston (2008) find that government transfers and family labor earnings are quantitatively important for insuring permanent shocks. They estimate that the pass-through rate of permanent income shocks to consumption is around 22 percent.

[38] Manso (2016) makes a related point but focuses on income risk rather than consumption risk, with the key source of insurance being outside options in paid-employment. If this is an important source of insurance, we should find differences in income risk distributions for the mostly switchers and the primarily self-employed, but we do not.

Figure 6: Self-Employment Exit Rates, 2001–2015



Notes: The figure shows the fraction of individuals whose status is self-employed in year $t-1$ and paid- or non-employed in year t. Age and time effects are extracted using the procedure outlined in Section 4.1.

we disaggregate these data further and ask how different subgroups of the self-employed population fared during the downturn.

Finding 8. IRS data show large declines in income growth during the Great Recession of 2008–2009 for self-employed in cyclically sensitive industries but no rise in exit rates.

In Table 8, we report estimates for the time effects $\Delta\beta_{g,t}$ relative to the 2008–2009 average group income in $t-1$ for a select set of cyclically sensitive subgroups of the population. The subgrouping is done by industry, education, and employment group. We selected industries in which we saw the largest income declines for the self-employed and then reported results by education and employment group. Not surprisingly, given the timing of the housing boom and bust, two cyclically important sectors are real estate and construction. Take, for example, those categorized as primarily self-employed. Average income growth in 2008–2009 for this group was -50 percent in real estate and -33 percent for construction. Less-educated owners fared somewhat better but still saw significant income declines. Individuals with fewer years in self-employment than the primarily self-employed also had lower average growth—in large part because many of them were in paid-employment in that period. The paid-employed had only modest declines, even in these highly cyclical sectors.

Despite huge declines in incomes for the self-employed during the Great Recession, we do not find an increase in exit rates. In Figure 6, we plot exit rates from self- employment into either non- or paid-employment over the 2001–2015 sample period. (In the online appendix, we report

Table 8: Estimated Time Effects Relative to Average Income (%), 2008–2009
Cyclically Sensitive Groups by Education and Industry

Education/ Industry	Primarily Self-employed	Mostly Switching	Primarily Paid-employed
College educated			
Real estate	−50	−18	−4
Construction	−33	−16	−4
Accomodation	−22	−12	−2
Manufacturing	−21	−14	−3
Retail trade	−14	−10	−3
Professional services	−7	−6	−1
Not college educated			
Manufacturing	−24	−9	−3
Construction	−15	−9	−5
Agriculture	−7	0	−1
Retail trade	−6	−4	0
Transportation	−4	−4	−2
Other services	−2	−2	0

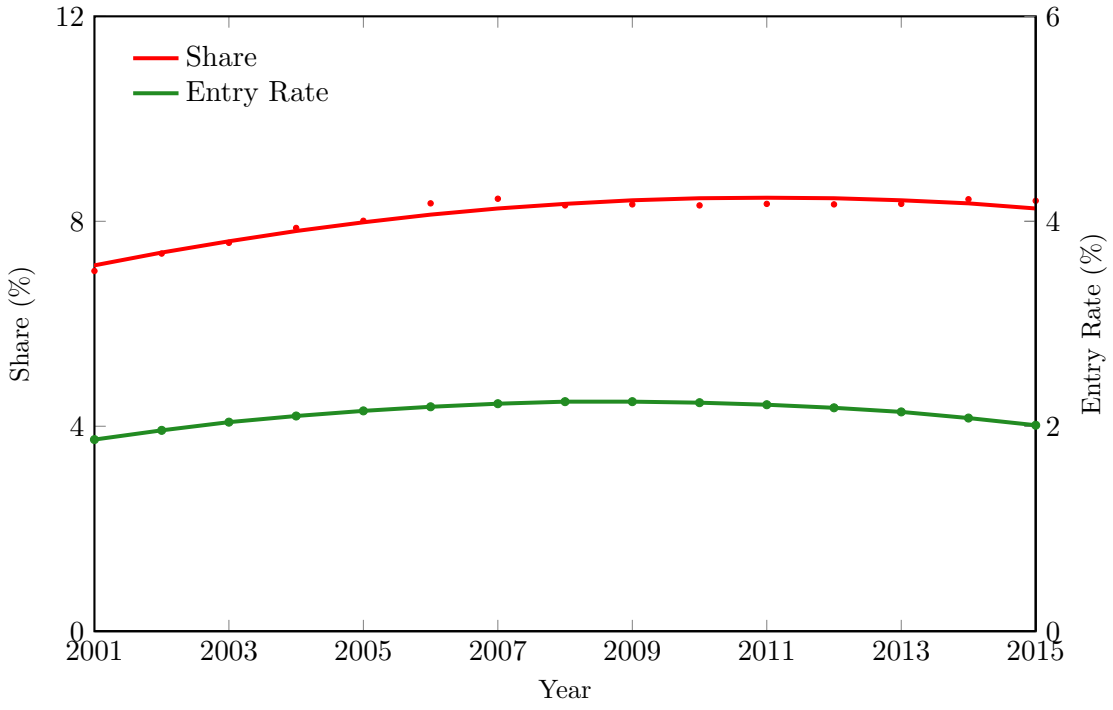
Notes: The table reports weighted averages of estimated time effects for select groups g at time t, that is, $\Delta\beta_{g,t}$, which are divided by average income $\overline{y}_{g,t}$. Weights are constructed from group counts.

exit rates—overall and separately into non- or paid-employment—by year and by age.) Because our population is aging over the sample, we extract age and time effects using the econometric procedure outlined in Section 4.1, both here and later when constructing entry rates and shares of the population that are self-employed.[39] The figure shows that the overall exit rate is nearly the same at the beginning and end of the sample, with no discernible change during the Great Depression. This pattern holds even for individuals working in cyclically sensitive industries: they experience large shocks to income but stay in business. The lack of cyclical variation around 2008–2009 suggests that self-employment was not used as a hedge against unemployment risk (see, for example, Alba-Ramirez (1994), Evans and Leighton (1989), Rissman (2003), and Rissman (2007)).

Thus far, we have investigated the life-cycle profiles of entrepreneurs and the benefits and risks of entrepreneurship, with a particular focus on individuals that are primarily employed and either tried self-employment or are primarily paid-employed over the sample. We turn next to consider the entry decision of individuals switching from paid- to self-employment and the potential impediments they face.

[39]To ensure that we capture secular trends in the data, we use a normalization that the time effect in the first year of the sample is set to the average rate in that year.

Figure 7: Self-Employment Shares and Entry Rates, 2000-2015



Notes: The figure shows two time series: (*i*) the share of the total population in IRS sample (C) of Table 1 categorized as self-employed using criteria (3) and (*ii*) the fraction of individuals whose status is paid- or non-employed in year $t-1$ and self-employed in year t. Age and time effects are extracted using the procedure outlined in Section 4.1.

5 Impediments to Entrepreneurship

A common narrative in the entrepreneurial choice literature is that entrants face significant impediments. For example, they lack financing, experience, or insurance—all of which are needed when starting a new business and successfully weathering volatile cash flows. In this section, we investigate this narrative by comparing entrepreneurial choices of different subgroups of the population. We start by documenting patterns of entry rates by age and time and then turn to the potential impediments to switching from paid- to self-employment. Finally, we track business founders to glean insights into impediments faced by owners in the early years of operating a new business.

5.1 Entry into Self-Employment

There has been an active literature on business "dynamism" in the U.S. economy following the work of Decker et al. (2014), who documented a secular decline in startup rates of employer firms using U.S. Census Bureau data. Several papers in this literature have used entrepreneurial choice models to understand the causes of declining business dynamism (see, for example, Salgado (2020), Jiang and Sohail (2023), and Kozeniauskas (2024)). They argue that better opportunities for high-skilled paid-employed workers due to skill-biased technological change are responsible for declining dynamism. This narrative is substantiated by survey-based evidence on declining entry rates into

self-employment, rising exit rates from self-employment, and an overall declining share of self-employed in the last few decades. Here, we use our administrative data to investigate this narrative.

Finding 9. IRS data show that entry rates for business owners and self-employment shares are not declining over the sample period.

In Figure 7, we report entry rates (left axis) and self-employment shares (right axis) by year after extracting age and time effects. We find that the entry rates by year are remarkably constant over time. In the online appendix, we provide additional evidence for owners with employees in light of U.S. Census Bureau reports of a decline in startup rates for employer firms, most notably during the Great Recession period. It turns out that entry rates for owners with employees show a modest decline over the sample, but there are no discernible changes in the Great Recession for either of the samples. Given that exit and entry rates have no trend in our sample, one would expect that the share of self-employed individuals is also stable over time. We verify this next by computing the fraction of our main sample population that is categorized as self-employed each year. As in the case of the entry and exit rates, we extract a time effect using the procedure in Section 4.1. In Figure 7, we show that the share of self-employed rises modestly between 2001 and 2006 and is then flat over rest of the sample period, most notably during 2008–2009.

While the evidence on entry rates provides a prima facie case that there are few impediments to entrepreneurship, we turn next to more direct evidence on whether potential entrants have sufficient liquidity, experience, and insurance.

5.2 Financing

A common starting point for discussions on impediments to self-employment entry is the issue of liquidity constraints for business startups. A standard narrative in the literature, such as that presented by Evans and Leighton (1989), posits that startup costs are high and that individuals often lack sufficient liquid wealth or pledgeable collateral to raise additional liquidity to finance them. This narrative is typically examined by exploring the correlation between self-employment entry and "wealth shocks," which are proxied by various instruments such as untimely deaths, lottery winnings, and differential exposures to housing or stock price growth.[40] In this section, we revisit this narrative using our administrative tax data in two ways. First, we examine cross-region differences in house price growth, as in Schmalz, Sraer, and Thesmar (2017). Second, we compare pre-entry liquid wealth measures for two groups of entrants with similar characteristics but different timing of entry.

Finding 10. Estimates based on IRS data suggest that appreciated housing values did not lead to an increase in entrepreneurial entry.

We follow the empirical strategy of Schmalz, Sraer, and Thesmar (2017), who use French administrative data, and we estimate the following equation:

$$\begin{aligned} E_{i,j,t+1} = \beta_0 &+ \beta_1 \cdot Owner_{i,t} \times \Delta p_{j,t-6\to t-1} + \beta_2 \cdot Owner_{it} \\ &+ \beta_3 \cdot Owner_{i,t} \times \Delta u_{j,t-6\to t-1} + \beta_4 \cdot \Delta u_{j,t-6\to t-1} \\ &+ \beta_5 \cdot Z_{i,t} + \beta_6 \cdot Z_{i,t} \times \Delta p_{j,t-6\to t-1} + \delta_{jt} + \epsilon_{i,j,t}, \end{aligned} \tag{14}$$

[40] The literature on this topic is extensive, so we cite a select few papers as examples. See Hurst and Lusardi (2004) for inheritances, Holtz-Eakin, Joulfaian, and Rosen (1994) for untimely deaths, Golosov et al. (2024) for lottery winnings, Schmalz, Sraer, and Thesmar (2017) for house price growth, and Ring (2023) and Chodorow-Reich et al. (2024) for stock price growth.

Table 9: Differences in Past Incomes for Switchers by Age, Current versus Future Switcher (Thousands, 2012$)

Past income type/ Percentile	Ages 30	40	50	60
Interest income				
25^{th}	−0.1	−0.4	−1.1	−1.8
50^{th}	0.0	−0.1	−0.2	−0.4
75^{th}	0.0	0.0	0.0	0.0
Dividend income				
25^{th}	−0.1	−0.3	−0.9	−1.6
50^{th}	0.0	0.0	−0.2	−0.3
75^{th}	0.0	0.0	0.0	0.0
Capital gains				
25^{th}	−0.2	−0.9	−3.6	−6.3
50^{th}	0.0	−0.1	−0.5	−1.1
75^{th}	0.0	0.0	0.0	0.2
Spousal wages				
25^{th}	−4.7	−14.6	−19.5	−20.3
50^{th}	−0.9	−2.4	−2.6	−1.8
75^{th}	0.0	0.0	0.0	0.0
Own wages				
25^{th}	−0.3	0.3	0.3	1.5
50^{th}	0.0	1.0	1.1	2.4
75^{th}	0.3	2.5	3.4	5.1

Notes: The table reports the time-averaged past incomes (in thousands of 2012 dollars) of current switchers less the time-averaged past incomes of peers that have similar characteristics but switch later. The sample used to construct the results includes individuals with at most one observed switch between paid- and self-employment. Differences in incomes are shown for selected percentile ranks and ages. To ensure that no confidential information is disclosed, reported percentiles are computed as an average of observations around the value listed in the table.

where $E_{i,j,t+1}$ is a dummy variable equal to 1 if individual i is living in state j in year t, has not been in self-employment over the last three years, and then switches to self-employment in year $t+1$; $Owner_{i,t}$ is equal to 1 if the individual owns a house in year $t-1$; $\Delta p_{j,t-6\to t-1}$ is the cumulative house price growth in state j between year $t-6$ and $t-1$; $\Delta u_{j,t-6\to t-1}$ is the change in the unemployment rate in state j between year $t-6$ and $t-1$; the vector $Z_{i,t}$ contains control variables; and δ_{jt} is a state by year fixed effect. The control variables in our case are: education, gender, the log of the previous year's income, industry, and age.[41] The coefficient of interest in

[41] Unlike Schmalz, Sraer, and Thesmar (2017), we do not include a foreign dummy or information about the job of the respondent's father.

equation (14) is β_1. Schmalz, Sraer, and Thesmar (2017) emphasize that the interaction of price appreciation and ownership rates—as opposed to just including price appreciation alone—is critical for their results.

We determine ownership status by analyzing deductions related to property taxes and mortgage interest payments on Schedule A of Form 1040. Individual i is labeled an "owner" if they have nonzero deductions for property tax or mortgage interest and a "full owner" if deductions are present only for property taxes. Following the approach of Schmalz, Sraer, and Thesmar (2017), we estimate (14) using both ownership measures. State-level house prices are measured using indices constructed by the Federal Housing Finance Agency, while state-level unemployment rates are obtained from the Bureau of Labor Statistics' local unemployment rate series.

We estimate equation (14) using OLS and construct the standard errors as in Schmalz, Sraer, and Thesmar (2017). When we include all homeowners, the coefficient β_1 in (14) is positive, with the estimate equal to 0.0012, and precisely estimated. However, the estimate is economically tiny: going from the 25^{th} to the 75^{th} percentile of house price growth (a 21 percentage point increase) leads to a 0.025 percentage point increase relative to the unconditional entry rate of 2 percent. When we restrict to full ownership—which is the more relevant treatment for this study—we find that the coefficient is negative and insignificantly different from zero. In other words, in contrast to Schmalz, Sraer, and Thesmar (2017), we do not find strong evidence that an increase in collateral values leads to a higher probability of becoming an entrepreneur.[42]

Finding 11. IRS data show that most entrants to self-employment have lower past asset incomes from liquid accounts than future entrants with similar characteristics.

Next, we test whether individuals who plan to start businesses need to accumulate liquid wealth to cover startup costs. This requires us to take a stand on how to measure pre-entry liquid wealth and to construct an appropriate comparison group for individuals transitioning to self-employment. To measure liquid wealth, we utilize information from Forms 1099-INT, 1099-DIV, and 1099-B, which allows us to categorize asset income into interest, dividends, and capital gains. According to an analysis of fixed-income asset returns by Smith, Zidar, and Zwick (2023), bank deposits are the primary source of interest income, particularly for individuals below the 98^{th} percentile of the adjusted gross income distribution. Smith, Zidar, and Zwick (2023) apply a uniform rate of return (which varies over time) across all bank deposits when capitalizing income. Based on this, our main proxy for liquid wealth is interest income. For robustness, we also consider broader measures of asset income.

Next, we define the comparison groups. The treated group consists of individuals for whom we observe a single transition from paid employment to self-employment, along with three years of pre-transition data. This group is compared to individuals who match on characteristics—namely, birth year, gender, industry, marital status, homeownership status, three years of lagged employment status (whether paid or non-employed), and the percentile of past wage income—but transition to self-employment at a later age.[43] For robustness, we also consider an expanded control group that includes non-switchers. Let $x_{i,t}$ be the proxy for pre-entry liquid wealth—say, past interest income—for individual i at time t, and let $x_{m(i),t}$ be the same variable for all matched peers $m(i)$ of individual i. Then we compute the difference Δ_{it} in the averages of variable x before

[42]The limited relevance of the collateral channel that we find is also echoed in the study by Kerr, Kerr, and Nanda (2022), who analyze a policy reform in Texas enabling home equity loans and lines of credit.

[43]In these comparisons, we use yearly indicators of marital status (married or not) rather than the time-invariant notion of being "mostly" married.

the switch as

$$\Delta_{it} = \frac{1}{3}\left[\sum_{j=1}^{3} x_{i,t-j} - \frac{1}{N_{m(i)}}\sum_{m(i)}\sum_{j=1}^{3} x_{m(i),t-j}\right]. \tag{15}$$

In the first three rows of Table 9, we report the interquartiles of the differences in past interest income by age of entry into self-employment. A positive value indicates current switchers have higher past income than future switchers. We find that most of our current switchers have *lower* interest income than similar peers switching later. If there were binding liquidity constraints and no other sources of liquidity, we would expect the opposite: individuals starting in the current period would have enough liquid wealth to pay the start up financing, whereas future switchers would be delayed in entry due to a lack of liquid wealth. While fixed-income assets are arguably the most relevant category for analyzing liquidity constraints, we find similar results when we compare other asset incomes—namely, dividends and capital gains—for the current and future switchers. We also find similar results when comparing spousal wages, which could be used to finance startup costs. From these comparisons, we conclude that most switchers are negatively selected on liquidity.

5.3 Experience

For first-time business owners, what may be needed even more than financial capital is human capital—for example, industry and managerial expertise. Consider the case of doctors working at a hospital to gain the necessary experience while in paid-employment before setting up their own practices. If this is indeed what happens, then it should be no surprise that individuals in our mostly switcher subsample—many of whom work in health care and professional services—have steeper income profiles than the paid-employed. This aligns with the depiction by Murphy, Shleifer, and Vishny (1991) of "superstar" entrepreneurs: the ablest individuals who achieve higher returns by establishing their own businesses and scaling them effectively. On the other hand, this portrayal of entrepreneurial selection is completely opposite of that based on survey evidence, most notably Evans and Leighton (1989), who find that individuals entering self-employment are not the ablest stars—lawyers, doctors, and engineers—but rather the low-wage "misfits" that are pushed into self-employment.[44] Here, we assess these starkly different views by repeating the analysis of Section 5.2 and comparing past labor income of current switchers with that of observationally similar peers that remained in paid-employment.[45]

Finding 12. IRS data show that most entrants to self-employment have higher past labor income than future entrants with similar characteristics.

In the last row of Table 9, we report the results. More specifically, we report the interquartiles of the difference in past income by age of switch. A positive value indicates switchers have higher past labor income than future switchers. We find that early switchers have similar past incomes to non-switchers, and over time the gap becomes larger and more favorable for the switchers. These findings hold up even if we focus exclusively on those in paid-employment prior to the switch. From this exercise, we conclude that most switchers are positively selected on past productivity.

5.4 Founders

In this section, we analyze *founders*—defined as S corporation and partnership owners for whom we can match establishment dates of business with the filing of Schedule K-1 in the year the business

[44] See also Alba-Ramirez (1994), Rissman (2003), and Rissman (2007).

[45] In this case, we do not condition on past labor income as we did in the asset analysis of Section 5.2.

Table 10: Year of First Positive Income for Founders of S Corporations and Partnerships

	Share, First Positive Income in Year:			
Income type	1	2	3	Total
Business net income	53.5	17.7	7.8	78.9
Owner total income	88.0	5.3	1.9	95.3
Adjusted gross income	95.2	2.0	0.7	97.8
Own wages	57.1	4.5	3.0	64.5
Spousal wages	55.2	4.9	2.8	62.9
Interest	74.7	5.9	2.9	83.5
Dividends	52.9	8.6	5.3	66.9
Capital gains	25.7	9.1	6.5	41.3

Notes: The table reports the share of founders that have their first positive income in years 1, 2, or 3. Income categories are listed in the first column. The sample of individuals used to compute these shares includes founders of S corporations and partnerships, who started their business in the same year the business is established and have at least eight years of tax filings with receipts or deductions.

is started. Analyzing founders and their ventures provides insights into the challenges faced by business owners during the early years of operation. For these young businesses, we examine their profitability, access to external financing, and the alternative income sources that help owners navigate initial losses.[46]

Finding 13. IRS data show that most businesses are not profitable in their first few years, but owners have enough income from other sources to ensure positive adjusted gross income on their individual tax filings.

We start by documenting the business losses in the first three years of operations. We find that 47 percent of businesses have losses in the first year, 37 percent in the second, and 34 percent in the third. Flipping this around, we can ask: when did these businesses first have a positive net income? The answers are shown in the first row of Table 10. The answer in the first column is something we already know: if 47 percent had no income or a loss, then 53 percent had a positive business net income. In the second year operating, 18 percent had their first profitable year, and in the third year, 8 percent more were in the black. By the fourth year, many owners have still not reported a positive net income.

From the business filings, we can also infer how many of the founders are using external debt financing, which could be potentially critical for young businesses that have losses in the first years of operation. For those that are borrowing, we use information on the ratio of interest deductions

[46] We restrict attention to businesses that have at least eight years of consecutive tax filings with business receipts or deductions in order to track them over time. We also analyzed founders across different cohorts but found negligible differences in any results.

to receipts to determine how important this source of funding is.[47] The deductions include interest expenses on credit cards, bank loans, government loans, mortgages, home equity loans, bonds, and any other debts relating to the business. It is possible that these businesses were denied bank financing, but one would expect that they would at least use credit cards or other forms of easy-to-access financing when desperate for funds, even if costs are higher. We find a large share of startups with no interest deductions at all, not even interest on credit card balances. For example, in the first year, 60 percent of businesses that had positive sales reported no interest expenses. If we rank businesses by the interest-to-sales ratio, we find that even at the 75^{th} percentile, the ratio is 0.005, which corresponds to a level of debt that is only 10 percent of sales, assuming an interest rate of 5 percent. Tracking these businesses over time, we find that external debt financing remains low for most businesses.[48]

Given that most of the businesses do not use external debt financing and have low or negative net incomes in the early years, a natural question to ask is: how did the founders fare? To answer this question, we compute several aggregated measures of income at the level of the owner. The results are shown in the second and third row of Table 10. We find that 88 percent have their first positive total income from either paid- or self-employment. Expanding the income measure to adjusted gross income, which includes own wages from paid-employment, spousal wages, asset incomes, and other incomes listed on Form 1040, we find that nearly all owners—roughly 95 percent—have positive adjusted gross income.

In other words, when starting a new business, owners rely on other sources of labor earnings, through either paid-employment or other business enterprises. Thus, even though most businesses have losses, few owners have negative individual incomes.

6 Conclusions

Much has been written about the nature of entrepreneurship, but our knowledge base is built up from analyses of very different and usually limited samples of individuals, which on the whole provide a narrative reminiscent of the parable of the blind men and an elephant. Each man learns about the elephant by touching only one part of the body, drawing conclusions that the elephant is like a wall, snake, spear, tree, fan, or rope, depending on what part they had touched. Analogously, the literature on entrepreneurship has an array of narratives, describing the typical business owner in many possible ways: as a gig worker seeking flexible arrangements, a misfit avoiding unemployment spells, an inventor seeking venture capital, a tax dodger misreporting income. To provide a more complete picture of the nature of entrepreneurship, we used U.S. administrative tax data to assemble a novel longitudinal database of business owners. Specifically, we analyzed patterns of income growth and determinants of entrepreneurial choice for a large population of business owners.

These data provide new insights into the central questions of the entrepreneurship literature and will hopefully prove useful for researchers interested in calibrating models of self-employment and business formation. Contrary to earlier studies based on surveys plagued by underrepresentation in the right tail of the income distribution, we find that non-pecuniary benefits of self-employment are not substantial when considering the source of most business income. Most entrepreneurs that persist in business have higher earnings growth than in paid employment. With insurance from the most adverse shocks, we find that self employment is an attractive option and not puzzling from

[47] Interest can be deducted even if personal property secures the loan, as long as it is an expense for the business. For our calculations, we use lines 1c and 13 on Form 1120-S for S corporations and lines 1c and 15 on Form 1065 for partnerships, respectively.

[48] These findings are consistent with the Kauffman Firm Survey that studies an eight-year panel of businesses founded in 2004.

a risk versus return perspective as previously thought. Consistent with this is our finding of no change in exit or entry rates during the Great Recession even though incomes of business owners in cyclically sensitive sectors fell dramatically. The longitudinal nature of our data also allowed us to revisit key questions about possible impediments to entry, such as a lack of financing, experience, or insurance. Contrary to survey-based evidence, we find little evidence of collateral or liquidity constraints or a lack of insurance, and we find that current entrants earn more in paid-employment prior to starting their business than similar peers who enter later.

We hope and expect that these findings will spur new research that will further enhance our understanding of the who, what, why, and how of entrepreneurial studies. Much more can be learned about their sources of insurance and financing as businesses grow and expand. More can be learned about the time investments owners make to grow their businesses. More can be learned about true incomes for the population, with the most important being employer benefits that are not included on tax forms and underreported incomes that should be. More can be done to link individual and business tax filings and provide a broader picture, especially for owners with multiple businesses.

7 Data Availability

Detailed information about the confidential tax data used in this article can be found in the Harvard Dataverse https://doi.org/10.7910/DVN/SSRRMX (Bhandari et al. (2025)).

References

Abowd, John M. and David Card. 1989. "On the Covariance Structure of Earnings and Hours Changes." *Econometrica* 57 (2):411–445.

Abraham, Katherine G., John C. Haltiwanger, Claire Hou, Kristin Sandusky, and James R. Spletzer. 2020. "Reconciling Survey and Administrative Measures of Self-Employment." *Journal of Labor Economics* 39 (4):825–860.

Alba-Ramirez, Alfonso. 1994. "Self-employment in the Midst of Unemployment: The Case of Spain and the United States." *Applied Economics* 26 (3):189–204.

Ameriks, John, Joseph Briggs, Andrew Caplin, Matthew D. Shapiro, and Christopher Tonetti. 2020. "Long-term-care Utility and Late-in-life Saving." *Journal of Political Economy* 128 (6):2375–2451.

Attanasio, Orazio P., James Banks, Costas Meghir, and Guglielmo Weber. 1999. "Humps and Bumps in Lifetime Consumption." *Journal of Business and Economic Statistics* 17 (1):22–35.

Auten, Gerald and Patrick Langetieg. 2023. "The Distribution of Underreported Income: What Can We Learn from the NRP?" Working paper, U.S. Department of Treasury.

Bhandari, Anmol, Tobey Kass, Thomas May, Ellen McGrattan, and Evan Schulz. 2025. "Replication Code for: 'On the Nature of Entrepreneurship'." `https://doi.org/10.7910/SSRRMX`. Harvard Dataverse.

Bhandari, Anmol, Paolo Martellini, and Ellen R. McGrattan. 2024. "A Theory of Business Transfers." Working paper, University of Minnesota.

Bhandari, Anmol and Ellen R. McGrattan. 2021. "Sweat Equity in U.S. Private Business." *Quarterly Journal of Economics* 136 (2):727–781.

Blundell, Richard, Luigi Pistaferri, and Ian Preston. 2008. "Consumption Inequality and Partial Insurance." *American Economic Review* 98 (5):1887–1921.

Bollinger, Christopher R., Barry T. Hirsch, Charles M. Hokayem, and James P. Ziliak. 2019. "Trouble in the Tails? What We Know about Earnings Nonresponse 30 Years after Lillard, Smith, and Welch." *Journal of Political Economy* 127 (5):2143–2185.

Cagetti, Marco and Mariacristina DeNardi. 2006. "Entrepreneurship, Frictions, and Wealth." *Journal of Political Economy* 114 (5):835–870.

Calvet, Laurent E., John Y. Campbell, Francisco Gomes, and Paolo Sodini. 2021. "The Cross-section of Household Preferences." Working paper, Harvard University.

Catherine, Sylvain. 2022. "Keeping Options Open: What Motivates Entrepreneurs?" *Journal of Financial Economics* 144 (1):1–21.

Chetty, Raj, Friedman John N., Emmanuel Saez, and Danny Yagan. 2018. "The SOI Databank: A Case Study in Leveraging Administrative Data in Support of Evidence-based Policymaking." *Statistical Journal of the IAOS* 34:99–103.

Chodorow-Reich, Gabriel, Plamen T. Nenov, Vitor Santos, and Alp Simsek. 2024. "Stock Market Wealth and Entrepreneurship." Working Paper 32643, NBER.

De Nardi, Mariacristina, Eric French, and John B. Jones. 2010. "Why Do the Elderly Save?" *Journal of Political Economy* 118 (1):39–75.

Deaton, Angus. 1997. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. World Bank.

DeBacker, Jason, Vasia Panousi, and Shanthi Ramnath. 2022. "A Risky Venture: Income Dynamics Among Pass-Through Business Owners." *American Economic Journal: Macroeconomics* 15 (1):444–474.

Decker, Ryan, John Haltiwanger, Ron Jarmin, and Javier Miranda. 2014. "The Role of Entrepreneurship in US Job Creation and Economic Dynamism." *Journal of Economic Perspectives* 28 (3):3–24.

Epstein, Larry G. and Stanley E. Zin. 1989. "Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework." *Econometrica* 57 (4):937–969.

Evans, David S. and Linda S. Leighton. 1989. "Some Empirical Aspects of Entrepreneurship." *American Economic Review* 79 (3):519–535.

Garin, Andrew, Emilie Jackson, and Dmitri Koustas. 2022. "New Gig Work or Changes in Reporting? Understanding Self-Employment Trends in Tax Data." Working Paper 2022–67, Becker Friedman Institute.

Golosov, Mikhail, Michael Graber, Magne Mogstad, and David Novgorodsky. 2024. "How Americans Respond to Idiosyncratic and Exogenous Changes in Household Wealth and Unearned Income." *Quarterly Journal of Economics* 139 (2):1321–1395.

Gourinchas, Pierre-Olivier and Jonathan A. Parker. 2002. "Consumption over the Life Cycle." *Econometrica* 70 (1):47–89.

Guvenen, Fatih, Fatih Karahan, Serdar Ozkan, and Jae Song. 2021. "What Do Data on Millions of U.S. Workers Reveal About Lifecycle Earnings Dynamics?" *Econometrica* 89 (5):2303–2339.

Guvenen, Fatih, Serdar Ozkan, and Jae Song. 2014. "The Nature of Countercyclical Income Risk." *Journal of Political Economy* 122 (3):621–660.

Hall, Robert E. 1968. "Technical Change and Capital from the Point of View of the Dual." *Review of Economic Studies* 35 (1):35–46.

Hall, Robert E and Susan E Woodward. 2010. "The Burden of the Nondiversifiable Risk of Entrepreneurship." *American Economic Review* 100 (3):1163–1194.

Hamilton, Barton H. 2000. "Does Entrepreneurship Pay? An Empirical Analysis of the Returns to Self-Employment." *Journal of Political Economy* 108 (3):604–631.

Holtz-Eakin, Douglas, David Joulfaian, and Harvey S. Rosen. 1994. "Entrepreneurial Decisions and Liquidity Constraints." *RAND Journal of Economics* 25 (2):334–347.

Hurst, Erik and Annamaria Lusardi. 2004. "Liquidity Constraints, Household Wealth, and Entrepreneurship." *Journal of Political Economy* 112 (2):319–347.

Hurst, Erik and Benjamin Wild Pugsley. 2011. "What Do Small Businesses Do?" *Brookings Papers on Economic Activity* 2011 (2):73–118.

Imboden, Christian, John Voorheis, and Caroline Weber. 2023. "Self-Employment Income Reporting on Surveys." CES Working Paper 23-19, U.S. Census Bureau.

Jiang, Helu and Faisal Sohail. 2023. "Skill-biased Entrepreneurial Decline." *Review of Economic Dynamics* 48 (2):18–44.

Johns, Andrew and Joel Slemrod. 2010. "The Distribution of Income Tax Noncompliance." *National Tax Journal* 63 (3):397–418.

Kartashova, Katya. 2014. "Private Equity Premium Puzzle Revisited." *American Economic Review* 104 (10):3297–3334.

Kerr, Sari Pekkala, William R. Kerr, and Ramana Nanda. 2022. "House Prices, Home Equity and Entrepreneurship: Evidence from U.S. Census Micro Data." *Journal of Monetary Economics* 130:103–119.

Kozeniauskas, Nicholas. 2024. "What's Driving the Decline in Entrepreneurship?" Working paper, Bank of Portugal.

Lazear, Edward P. and Robert L. Moore. 1984. "Incentives, Productivity, and Labor Contracts." *Quarterly Journal of Economics* 99 (2):275–296.

Levine, Ross and Yona Rubinstein. 2017. "Smart and Illicit: Who Becomes an Entrepreneur and Do They Earn More?" *Quarterly Journal of Economics* 132 (2):963–1018.

Lim, Katherine, Alicia Miller, Max Risch, and Eleanor Wilking. 2019. "Independent Contractors in the U.S.: New Trends from 15 Years of Administrative Tax Data." Soi working paper, Internal Revenue Service.

Lise, Jeremy and Fabien Postel-Vinay. 2020. "Multidimensional Skills, Sorting, and Human Capital Accumulation." *American Economic Review* 110 (8):2328–2376.

Manso, Gustavo. 2016. "Experimentation and the Returns to Entrepreneurship." *Review of Financial Studies* 29 (9):2319–2340.

Moskowitz, Tobias J. and Annette Vissing-Jorgensen. 2002. "The Returns to Entrepreneurial Investment: A Private Equity Premium Puzzle?" *American Economic Review* 92 (4):745–778.

Murphy, Kevin M., Andrei Shleifer, and Robert W. Vishny. 1991. "The Allocation of Talent: Implications for Growth." *Quarterly Journal of Economics* 106 (2):503–530.

Parker, Simon C. 2018. *The Economics of Entrepreneurship*. Cambridge University Press, 2 ed.

Ring, Marius A.K. 2023. "Entrepreneurial Wealth and Employment: Tracing Out the Effects of a Stock Market Crash." *Journal of Finance* 78 (6):3343–3386.

Rissman, Ellen. 2003. "Self-Employment as an Alternative to Unemployment." Working Paper 2003–34, Federal Reserve Bank of Chicago.

———. 2007. "Labor Market Transitions and Self-Employment." Working Paper 2007–14, Federal Reserve Bank of Chicago.

Salgado, Sergio. 2020. "Technical Change and Entrepreneurship." Working paper, Wharton.

Schmalz, Martin C., David A. Sraer, and David Thesmar. 2017. "Housing Collateral and Entrepreneurship." *Journal of Finance* 72 (1):99–132.

Smith, Matthew, Danny Yagan, Owen Zidar, and Eric Zwick. 2019. "Capitalists in the Twenty-First Century." *Quarterly Journal of Economics* 134 (4):1675–1745.

Smith, Matthew, Owen Zidar, and Eric Zwick. 2023. "Top Wealth in America: New Estimates Under Heterogeneous Returns." *Quarterly Journal of Economics* 138 (1):515–573.