Sources of U.S. Wealth Inequality: Past, Present, and Future

Joachim Hubmer, Per Krusell, and Anthony A. Smith, Jr.*

April 1, 2020

Abstract

This paper employs a benchmark heterogeneous-agent macroeconomic model to examine a number of plausible drivers of the rise in wealth inequality in the U.S. over the last forty years. We find that the significant drop in tax progressivity starting in the late 1970s is the most important driver of the increase in wealth inequality since then. The sharp observed increases in earnings inequality and the falling labor share over the recent decades fall far short of accounting for the data. The model can also account for the dynamics of wealth inequality over the period—in particular the observed U-shape—and here the observed variations in asset returns are key. Returns on assets matter because portfolios of households differ systematically both across and within wealth groups, a feature in our model that also helps us to match, quantitatively, a key long-run feature of wealth and earnings distributions: the former is much more highly concentrated than the latter.

1 Introduction

The distribution of wealth in most countries for which there is reliable data is strikingly uneven. There is also recent work suggesting that the wealth distribution has undergone significant movements over time, most recently with a large upward swing in dispersion in several Anglo-Saxon countries.[1] For example, according to the estimates in Saez & Zucman (2016) for the United States, the share of overall wealth held by the top 1% has increased from around 25% in 1980 to over 40% today; for the top 0.1% it has increased from less than 10% to over 20% over the same time period.

The observed developments have generated strong reactions across the political spectrum. In his 2014 book, *Capital in the Twenty-First Century*, Piketty is obviously motivated by the growing inequality in itself, but he also suggests that further increases in wealth concentration may lead to both economic and democratic instability. Conservatives in the U.S. have expressed worries as well: is the American Dream really still alive, or might it be that a large fraction of the population simply will no longer be able to

[*] The authors' affiliations are, respectively, University of Pennsylvania; Institute for International Economic Studies, NBER, and CEPR; and Yale University and NBER. For helpful comments the authors would like to thank Chris Carroll, Paolo Sodini, Harald Uhlig, and seminar participants at the 2015 SED Meetings, the 2015 Hydra Workshop on Dynamic Macroeconomics, the Seventh Meeting of the Society for the Study of Economic Inequality, the 2017 NBER Summer Institute, the 2019 NTA Conference, Bern, ECB, Johns Hopkins, Indiana, M.I.T., Oslo, Penn State, University of Pennsylvania, SOFI, and Yale. A previous version of this paper circulated as NBER Working Paper #23011 under the title "The Historical Evolution of the Wealth Distribution: A Quantitative-Theoretic Investigation".

[1] See, e.g., Piketty (2014) and Saez & Zucman (2016).

productively contribute to society? Given, for example, that parental wealth and well-being are important determinants of children's human capital accumulation, these are legitimate concerns regardless of one's political views. These concerns, moreover, have stimulated the proposal and discussion of a number of possible changes in policy. The primary aim of the present paper is, instead, to understand the determinants of the observed movements in wealth inequality. This aim is basic but well-motivated in light of the policy discussion: to compare different policy actions, we need a framework for thinking about what causes inequality and for addressing how any particular policy influences not only inequality but also other macroeconomic variables.

In an effort to understand the movements in wealth inequality, Piketty (2014) and its online appendix suggest specific mathematical theories and as part of the present study we examine those theories.[2] Our aim, however, is to depart instead from a more general, and by now rather standard, quantitative theory used in the heterogeneous-agent literature within macroeconomics: the Bewley-Huggett-Aiyagari model. This is a very natural setting for the study of inequality. This model incorporates rich detail at the household level along the lines of the applied work in the consumption literature, allowing several sources of heterogeneity among consumers. It is based on incomplete markets and, hence, does not feature the "infinite elasticity of capital supply" of dynastic models with complete markets.[3] This model also involves equilibrium interaction: inequality is determined not only by the individual households' reactions to changes in the economic environment in which they operate but also by their interaction, such as in the equilibrium formation of wages and interest rates, two key prices determining the returns to labor and wealth, respectively. Our aim is to see to what extent a reasonably calibrated model can account for the movements in wealth inequality since the mid-1960s as a function of a number of drivers, the importance of each of which we then evaluate in separate counterfactuals.[4]

We build on the model studied in Aiyagari (1994), i.e., we use the core setting of the recent literature on heterogenous agents in macroeconomics.[5] This kind of theoretical model is quantitative in nature: it is constructed as an aggregate version of the applied work on consumption. Moreover, inequality plays a central role in this model. We calibrate some key parameters of this model to match the wealth and income distributions in the United States in the mid-1960s and treat these distributions as representing a long-run steady state. In the 1960s, too, the dispersion of wealth was striking, and it is not immediate how to make the basic model match the data in this respect. In particular, the benchmark models in the literature do not readily produce long-run wealth inequality that is as striking as that observed: they do not produce wealth dispersion that goes much beyond earnings dispersion. The data shows, again wherever reliable data is available, a wealth Gini much above 0.5 (say, 0.8), whereas the earnings Gini is typically significantly below 0.5. In this paper we depart from the benchmark model by introducing portfolio heterogeneity across and within wealth groups. As we shall discuss in detail below, such heterogeneity

[2] The appendix is available here: http://piketty.pse.ens.fr/files/capital21c/en/Piketty2014TechnicalAppendix.pdf. See also Piketty (1995) and Piketty (1997) which develop theories of the dynamics of the wealth distribution.

[3] This elasticity refers to the long-run response of a household's savings to a change in the interest rate: in particular, with infinitely-lived consumers and complete markets the equilibrium interest rate is pinned down by the rate of time preference.

[4] We do not specifically study Piketty's "Second Fundamental Law", which is not a theory about inequality per se but about the aggregate capital-output ratio and which has also been extensively examined in Krusell & Smith (2015).

[5] The first application in this literature was one to asset pricing (the risk-free rate): Huggett (1993). Aiyagari (1994) addresses the long-run level of precautionary saving, whereas Krusell & Smith (1998) look at business cycles.

has recently surfaced as a striking feature of households' investment patterns. In particular, register data in Norway and Sweden (see Fagereng, Guiso, Malacrino, & Pistaferri (2020) and Bach, Calvet, & Sodini (2019)) have revealed, first, an average return that is increasing in the household's overall level of wealth; and, second, an idiosyncratic return component (because different households hold different types of assets) whose variance is also increasing in wealth.

Our first major finding is that, once portfolio heterogeneity, calibrated to the findings in Bach et al. (2019), is incorporated into the model, we replicate wealth inequality of the magnitude we see in the data. Thus, in order to match the agglomeration at the top, we do not need to consider discount-factor heterogeneity, as in Krusell & Smith (1998), or other mechanisms that raise the saving of the wealthiest.[6] Our model, which is fully nonlinear with household decision rules for saving whose slopes differ widely between the poorest and the richest, delivers a law of motion for wealth that becomes approximately linear in wealth for high wealth levels, with a random coefficient. It can thus be viewed as a microfoundation for the kind of models entertained in Piketty & Zucman (2015) (who simply assume linear laws of motion for wealth accumulation and either random saving propensities or random returns). A closely related setting is that in Benhabib, Bisin, & Luo (2019). These models, and by extension ours, generate a wealth distribution whose right tail is Pareto-shaped, a prominent feature in the data; we discuss this finding, and the relation to a number of other papers building on the same kind of reduced form, in detail in the paper.

With the resulting, realistic, starting wealth distribution, we then examine a number of potential drivers of wealth inequality over the subsequent period. One is tax rates: beginning around 1980 tax rates fell significantly for top incomes, so that tax progressivity in particular fell substantially. Thus, higher returns to saving in the upper brackets since that time can potentially explain increased wealth gaps between the rich and the poor. Another potential explanation for increased wealth inequality is the rather striking increases in wage/earnings inequality witnessed since the mid-1970s. Since at least Katz & Murphy (1992) it has been well-documented that the education skill premium has risen. Moreover, numerous studies have since documented that the premia associated with other measures of skill have also risen, as have measures of residual, or frictional, wage dispersion.[7] In terms of the very highest earners, Piketty & Saez (2003) document significant movements toward thicker tails in the upper parts of the distribution. So to the extent that this increased income inequality has translated into savings and wealth inequality, it could explain some of the changes we set out to analyze. Moreover, and very importantly as it will turn out, we feed in fluctuations in asset returns like those observed in the U.S. and that, given the systematic portfolio heterogeneity across wealth groups, may imply dynamic movements of wealth inequality. Finally, the share of total income paid to capital has increased recently, potentially contributing to increased wealth inequality (see, e.g., Karabarbounis & Neiman (2014b)). We consider this factor as well in this study.

[6] In our benchmark model we do in the end incorporate heterogeneity in discount factors, in part because the cross-sectional variance of returns for the wealthiest is so large that the shape of the right tail of the wealth distribution is, in fact, thicker than in the data. As explained in Section 6.5, we therefore adjust this variance and compensate by introducing a small amount of discount-factor heterogeneity. Although we do not explicitly calibrate our model to it, the empirical micro literature provides abundant support for such heterogeneity; see e.g. Cronqvist & Siegel (2015).

[7] See, e.g., Acemoglu (2002), Hornstein, Krusell, & Violante (2005), and Quadrini & Rios-Rull (2015).

Thus, the overall methodology we follow is to attempt to quantify the mechanisms just mentioned and then to examine their individual (and joint) effects on the evolution of wealth inequality from the 1960s. For the time period considered, we find, first, that the benchmark model does account well for the net increase in wealth inequality over the period. The model is more or less successful depending on what aspect of the wealth distribution is in focus. The shares of wealth held by the top 10%, the top 1%, and the top 0.1% exhibit net increases that are very similar in the model and in the data, though for the top 0.01% the benchmark model does not deliver enough of an increase. For the bottom 50%, the model's fit is also good. Second, in terms of the dynamics, the model also proves to be successful in replicating the marked U-shape of wealth inequality. Furthermore, the model delivers a time path for the ratio of capital to net output that is similar to the one in the data.

Turning to which specific features explain the largest fractions of the increase in wealth inequality, the marked decrease in tax progressivity is by far the most powerful force for the cumulative increase in wealth inequality.[8] First, other things equal, decreasing tax progressivity spreads out the distribution of after-tax resources available for consumption and saving. Second, decreasing tax progressivity increases the returns on savings, leading to higher wealth accumulation, especially among the rich for whom wages (earnings) are a smaller part of wealth. As for the dynamics, here swings in the returns of the different asset groups turn out to be crucial. In agreement with Kuhn, Schularick, & Steins (2019), we find that without portfolio heterogeneity, and without asset-price movements, we would not be able to understand the short- and medium-run movements in wealth inequality.

Wage inequality, on the other hand, has less clearcut effects on wealth. As we argue in our paper, it can both increase and decrease wealth inequality, depending on the nature of the increased earnings risk and on what wealth-inequality statistics one looks at. In some aggregate sense—measured by the shares of wealth held by the richest households—the kinds of earnings inequality we feed in on net contributes *negatively* to wealth inequality, taken together. We consider increases in earnings inequality of different kinds. We follow Heathcote, Storesletten, & Violante (2010) in modeling increased wage inequality as an increase in the riskiness of wage realizations around a mean. In a standard additive permanent-plus-transitory model of wages, we use the estimated time series in Heathcote et al. (2010) for the variances of the permanent and transitory shocks to wages. Both of those variances have increased over time, leading to a reduction in the share of wealth held by the richest for two reasons. First, increasing wage risk dampens the tendency of heterogeneity in returns or discount rates to drive apart the distribution of wealth.[9] In particular, as wage risk increases, poorer and less patient consumers—who are less well-insured against this risk through their own savings—engage in additional precautionary saving, compressing the distribution of wealth at the low end. Second, with more risk aggregate precautionary savings increase, reducing the equilibrium interest rate and reducing the relative wealth accumulation of the rich, for whom wage risk is also not so important. In sum, the increasing riskiness of wages compresses the wealth distribution at both ends.[10] At the same time, these increases in earnings risk do induce higher inequality if one looks at the dispersion

[8] These conclusions are in line with two studies of France and the U.S.: Piketty (2003) and Piketty & Saez (2003).

[9] As Becker (1980) shows, if discount rates are permanently different and there is no wage risk at all, then in the long-run steady state the most patient consumer owns all of the economy's wealth.

[10] Similar forces are at play in Krusell, Mukoyama, Şahin, & Smith (2009), but in the opposite direction: they find that reductions in wage risk that accompany the elimination of business cycles lead to higher wealth inequality.

of wealth within the bottom part of the distribution rather than within the whole distribution.

In addition, we follow Piketty & Saez (2003) by adding a Pareto-shaped tail to the wage distribution so as to match the concentration of earnings at the top of the earning distribution; the standard wage process (as in Heathcote et al. (2010)) does not match this extreme right tail well. Moreover, the right tail has thickened over this period, and accordingly we model this thickening as a gradually decreasing Pareto coefficient, based on the estimates in Piketty & Saez (2003). This element of increased wage inequality does generate more wealth inequality—because it occurs in a segment of the population where most workers are already rather well-insured through their own savings—but it is not so potent as to produce a net overall increase in wealth inequality from higher wage inequality. To allow for an increasing capital share over time we conduct an experiment using a CES production function with a somewhat higher than unitary elasticity between capital and labor. The resulting paths in this experiment differ only marginally from the case with unitary elasticity.

Given the role of portfolio heterogeneity and of asset-price movements, it is important to think more about the origins of these observations. In the present paper we take short-cuts in both these respects. First, we simply hard-wire the portfolio heterogeneity. The consumer making a saving decision knows, given the current level of wealth, what the return characteristics are (but has no choice but to accept them, i.e., cannot switch to holding different asset shares) and what they will be like henceforth. Since there is a higher average return as a function of wealth, the household therefore factors in this small amount of "increasing returns" to saving in setting the current saving rate. Interestingly, the household's choice of a saving rate is not very sensitive to the return characteristics, and hence a Solow-like constant saving rate comes close to approximating optimal behavior.[11] In particular, a model with myopic forecasts delivers very similar behavior to that in our benchmark (where agents have perfect foresight). Second, we do not attempt to solve for asset prices by clearing markes for each asset class. This would necessitate taking a stand on how to solve the equity premium puzzle and, more than that, also match returns for other asset classes—we incorporate houses and private equity as well, which are very important for the average household and the richest, respectively. The two shortcuts we take seem necessary at this stage; rather, we view our present paper as an important step forward in noting just how important portfolios and asset prices are for inequality. Taking the whole step forward in explaining them is one or two orders of magnitudes more challenging, but these steps definitely seem worth taking now.

What are the implications of our dynamic model of wealth inequality for the future? Quite strikingly, if the progressivity of taxes remains at today's historically low level, then wealth inequality will continue to climb and reach very high levels by, say, 2100: the top 10% will have an additional 10% of all of wealth, while the top 1% share will increase by more than 20%. Thus, decreasing the progressivity of taxes is a rather powerful mechanism for wealth concentration.

Our paper begins in Section 2 with a brief literature review, the purpose of which is to put our modeling in a historical perspective. We discuss the data on wealth inequality and its recent trends in Section 3. We describe the basic model in Section 4 and the implied behavior of the very richest in Section 5. Section 6 discusses the calibration in detail and Sections 7 and 8 the benchmark results for long-run

[11] Bach et al. (2019) document striking "stickiness" in individuals' portfolio choices. This is consistent with our saving rates being quite insensitive to the return characteristics.

wealth inequality and its historical evolution, respectively. A number of extensions are then included in Section 9. We conclude our paper in Section 10 with a brief discussion of potential other candidate explanations behind the increased wealth inequality and, hence, of possible future avenues for research.

2 Connections to the recent macro-inequality literature

The study of inequality in wealth using structural macroeconomic modeling can be said to have started with Bewley (undated), though in Bewley's paper the focus was not on inequality per se.[12] Bewley's paper was not completed—it stops abruptly in the middle—and the first papers to provide a complete analysis of frameworks like his are Huggett (1993) and Aiyagari (1994). A defining characteristic of these models is that long-run household wealth responds smoothly to the interest rate, so long as the interest rate is not too high (higher than the discount rate in the case without growth).

In their early papers, neither Bewley nor Huggett nor Aiyagari focused on inequality per se but rather on other phenomena related to inequality (asset pricing and aggregate precautionary saving in the latter two cases, respectively). Soon after, however, the macroeconomic literature that arose from these analyses began to address inequality directly. There were several reasons for this development. One was the interest in building macroeconomic models with microeconomic foundations in which heterogeneity could influence aggregates, i.e., cases that depart from the typical permanent-income behavior that characterizes the complete-markets model.[13] Another was an interest in wealth inequality per se and the challenge it posed: the difficulty that these models have in generating significant equilibrium wealth inequality. The difficulty is apparent in Aiyagari (1994), where the wage process is calibrated to PSID data (as an AR(1) in logs): the resulting wealth distribution is slightly more skewed than the wage distribution the model uses as an input, but not by much. The Gini index for wealth, in the stationary distribution of Aiyagari's model, is only around 0.4, whereas it is around 0.8 in the data. The purpose here is not to go over the entire literature aiming at matching the wealth distribution but several different extensions of the model have been proposed in order to match the data better. On some general level, successful paths forward involve introducing "more heterogeneity": typically in preferences (such as discount factors, as in Krusell & Smith (1998)), in the wage/earnings process (as in Castañeda, Días-Giménez, & Ríos-Rull (2003)), or in occupation (as in Cagetti & De Nardi (2006) or Quadrini (2000)).

More recently, a literature evolved that focuses on explaining the observed Pareto tail at the top of the wealth distribution. Benhabib, Bisin, & Zhu (2011) show analytically that the stationary wealth distribution in an overlapping-generations (OLG) economy with idiosyncratic capital return risk has a Pareto tail. Analogously, they provide analytical results for an infinite-horizon economy (Benhabib, Bisin, & Zhu, 2015). In Benhabib et al. (2019), they conduct a quantitative investigation of social mobility and the wealth distribution in an OLG economy with idiosyncratic returns, which are fixed over a life-time. In

[12] This model is of course not the first one with theoretical implications for inequality. An early example is Stiglitz (1969) who, building on his 1966 Ph.D. dissertation, studies the dynamics of the distributions of income and wealth in a neoclassical growth model with exogenous linear savings functions. A defining characteristic of the literature in focus here is that consumers face problems much like those studied in the applied consumption literature: they are risk-averse and choose optimal saving in the presence of earnings shocks for which there is not a full set of state-contingent markets.

[13] See, e.g., Krusell & Smith (1998) and Guerrieri & Lorenzoni (2017) for this line of work.

a stylized model, Gabaix, Lasry, Lions, & Moll (2016) demonstrate that the random growth mechanism that can generate the Pareto tail in the wealth distribution (either through idiosyncratic capital return risk or random discount factors) implies very slow transitional dynamics. Furthermore, Nirei & Aoki (2016) consider a stationary Bewley economy with investment risk.[14] In that setting they find that decreasing top tax rates can explain the increasing concentration of wealth at the top.

Most of the literature on Bewley models has considered only the stationary (long-run) wealth distribution. A recent exception is Kaymak & Poschke (2016), who in line with our analysis here aim to quantify the contributions of changes in taxes and transfers and in the earnings distribution to changes in the U.S. wealth distribution; we compare their results to ours in more detail below. Another recent paper of this sort is Aoki & Nirei (2017), which studies how a one-time drop in tax rates affects transitional dynamics in a setting with investment risk.

The present paper has three main characteristics that distinguish it from the just-discussed earlier work. The first characteristic is that, in contrast to all but a handful of studies, it addresses the *long-run as well as short- and medium-run* determinants of the wealth distribution. Second, our model is rather *comprehensive, in two ways*: (i) it considers all the main mechanisms that the literature discusses regarding the buildup in inequality and (ii) it looks at the full distribution of wealth, i.e., both the upper tail as well as at the bulk of the distribution. Our model generates a Pareto tail endogenously, because it delivers approximately linear saving dynamics for households—with a stochastic coefficient on wealth—as wealth grows large. The key measure of the fatness of the right wealth tail is the (inverse of the) Pareto coefficient. In the data, its value, as we elaborate on below and is also emphasized elsewhere, is significantly higher than that for the earnings distribution.[15] A model with earnings risk only will either not deliver a Pareto tail for wealth at all or, if earnings risk is itself Pareto, will deliver a Pareto tail for wealth of the same shape as for earnings.[16] To us, thus, stochastic returns to saving and/or stochastic discounting, which do deliver the correct right-tail shape of wealth, are essential for understanding the right tail of the wealth distribution in the long run. This sets our paper apart from other Aiyagari-based models. This includes Kaymak & Poschke (2016), which delivers a very nice account of the medium-run features of the bulk of the wealth distribution but which does not have its focus on, and does not fully account for, its right tail.[17] We have in common with Kaymak & Poschke (2016) that we also include a thorough discussion of the the model's predictions for the middle and lower parts of the wealth distribution. We discuss how our transitional results differ from theirs in detail in Section 8 below.

The third characteristic that sets our paper apart from, we believe, all of the above-mentioned literature and hence is the most novel, is that it incorporates *portfolio behavior that differs across households.* Wealthy households have portfolios with more risk and higher average return. In addition, there is a non-negligible idiosyncratic return component at all wealth levels, with an accentuation for the wealth-

[14] See also Toda (2014), which also studies a stationary economy with investment risk, and Toda (2018), which studies a Huggett-like economy with random discount factors.

[15] For an illuminating recent discussion, see Benhabib, Bisin, & Luo (2017).

[16] See Stachurski & Toda (2019).

[17] In Kaymak and Poschke's work, the long-run wealth distribution does not have a Pareto tail. Moreover, the fraction held by the top group in their study is as high as in the data only because the earnings inequality is assumed to be more extreme than what the available micro data suggests.

iest. These features are not free parameters in our model: we calibrate them to available micro data and, in particular, track the returns, by asset subgroup, over time. Because of the systematic differences in portfolio compositions and in the return to different portfolios over the period, we obtain predictions for the evolution of the wealth distribution and it turns out that this allows us to match the short- and medium-run dynamics surprisingly well. In particular, there is a marked U-shape of the top wealth shares over the time period under study, and none of the other papers in the literature can generate this shape. We conclude that return heterogeneity—in particular, both the systematically different portfolios across wealth levels (which are important for wealth inequality dynamics) and the stochastic idiosyncratic component (which is important for understanding the right tail of the long-run wealth distribution)—is central to an understanding of wealth inequality and its evolution over time. We therefore now consider it crucial in this area to turn our attention toward understanding the deep determinants of all these features of observed portfolio decisions.

A final relevant literature connection is that to Piketty's $r-g$ theory: our framework can be interpreted as giving support to an elaborate version of this theory. The elaboration involves (i) negligible emphasis on g; (ii) the interpretation of r as net of taxes; and (iii) the (crucial) recognition that r is heterogeneous across households and systematically different for different wealth levels, both because taxation is progressive and because portfolios are heterogeneous. It must be emphasized, however, that this theory primarily works for the right tail of the wealth distribution; for understanding the rest, the kind of analysis pursued by Kaymak & Poschke (2016) as well as that herein, seems necessary.

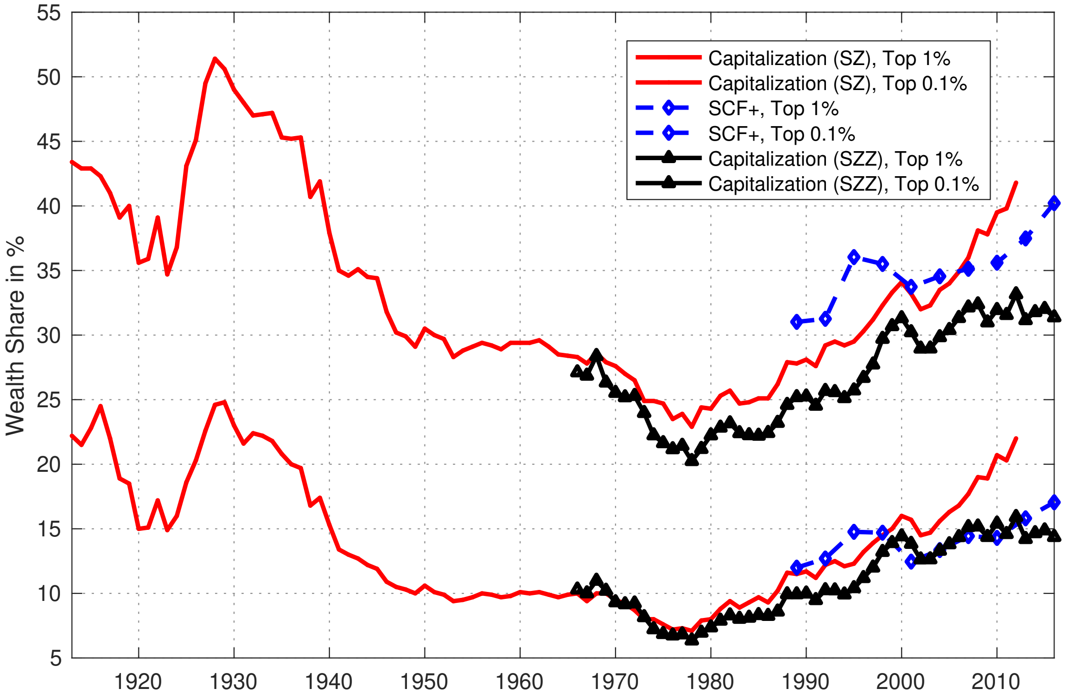
3 Measuring wealth inequality over time

Over the last century, the distribution of wealth in the United States has undergone drastic changes and we very briefly review data from some key studies here. Throughout the time period considered, wealth was heavily concentrated at the top. Figure 1 shows the evolution of the share of total wealth held by the top 1% and the top 0.1%, as measured using different estimation methods.

Considering all three methods jointly, top wealth inequality exhibits a U-shaped pattern in the twentieth century. At the same time, the magnitude of the increase in wealth concentration in the last thirty years differs substantially among estimation methodologies. We will calibrate the initial steady state of our model to the wealth shares estimated by Saez & Zucman (2016) and consequently compare the model transition to their estimates. Their estimates are especially useful for us as they allow for considering a group as small as the top 0.01%. Furthermore, they cover a long time period.

While the capitalization method that they use to back out wealth estimates does not suffer from the shortcomings of the SCF data (such as concerns about response-rate bias and exclusion of the Forbes 400), it is an indirect way of measuring wealth and as such has other drawbacks. For example, the tax data allows only for a coarse partitioning of capital income in asset classes and within each class returns are effectively assumed to be homogeneous. Since recent evidence based on both Norwegian and Swedish data (Fagereng et al. (2020) and Bach et al. (2019), respectively) shows significantly higher returns for high-wealth groups, the basic capitalization method suggests an over-prediction of wealth levels for the richest

Figure 1: Top wealth share measurements over time



The lines labelled "Capitalization (SZ)" display findings from Saez & Zucman (2016), who back out the stock of wealth held by a tax unit from observed capital income tax data. The lines labelled "Capitalization (SZZ)" display findings from Smith et al. (2019), who adjust the capitalization method for return heterogeneity within asset classes. The lines labelled "SCF+" refer to data from the Survey of Consumer Finances as reported by Smith, Zidar & Zwick, augmented with information on the Forbes 400, which are by design excluded from the SCF.

group. Indeed, Smith, Zidar, & Zwick (2019) adjust the method of Saez and Zucman for heterogeneity within asset classes and find a smaller increase in wealth concentration that is comparable to survey data. Therefore, we will in addition contrast our findings to their estimates.[18]

Another takeaway from Figure 1 is that the wealth distribution was quite stable in the 1950s and 1960s. In addition, some of the time series estimates we feed into our model start in 1967; we therefore take this year as the initial steady state in our model.

4 Model framework

What are the determinants of long-run wealth inequality, and what affects its dynamics? The present paper puts particular emphasis on these dynamics, but in order to understand them one also needs to take a stand on the longer-run drivers of wealth inequality. In particular, the framework we use for analyzing

[18] See also Kopczuk (2015) and Bricker, Henriques, Krimmel, & Sabelhaus (2016) for a detailed comparison of different measurement methods. In addition to the series presented here, Kopczuk & Saez (2004) use observed estate tax data to make inferences about the distribution of wealth. The resulting top wealth share series roughly confirm the Saez & Zucman (2016) findings until about 1980; afterwards, there is hardly any increase. This discrepancy has been linked to differential evolution of mortality rates across wealth groups (Saez & Zucman, 2019). Furthermore, Piketty, Saez, & Zucman (2018) and Saez & Zucman (2019) present revised capitalization method estimates that show a slightly smaller but comparable increase relative to Saez & Zucman (2016).

long-run inequality has important implications for dynamics, as we shall explain. As a background, let us first—in Section 4.1—very briefly recall some basic predictions for equilibrium wealth inequality from a set of standard models. In the subsequent sections, we will draw on these insights when formulating and interpreting the specific model we employ in our paper.

4.1 Long-run wealth inequality: a primer

Let us focus mostly on the predictions for inequality using dynastic models, i.e., frameworks where agents put value on their offspring and are altruistic in that respect. At the very end, we will briefly make comments on alternative assumptions in this regard. We will, for simplicity, also abstract from age dependence of either preferences or income streams and simply regard household i's present-value utility as being $\mathbb{E}_0 \sum_{t=0}^{\infty} \beta_i^t u_i(c_{it})$ and its income stream as a stationary process. Let us also consider a neoclassical production function $F(K_t, L)$, no technological change, and geometric depreciation of capital at rate δ.[19] That is, we have a standard optimizing growth model with more than one agent.

The permanent-income model Let us first consider a constant endowment stream. The consumer's budget constraint in our simplest setting is then $c_{it} + k_{i,t+1} = \omega_i w_t + (1 + r_t)k_{it}$, where w_t and R_t are the marginal products of labor and capital based on $F(K_t, L)$, and $r_t = R_t - \delta$; ω_i is agent i's endowment of labor in efficiency units. Let us also for illustration consider only two kinds of agents, A and B, with masses μ_A and μ_B, respectively. The key observation here is that if $\beta_A = \beta_B$, then any wealth distribution (k_A, k_B) is a steady state, so long as $\mu_A k_A + \mu_B k_B = K^\star$, where $K^\star$ satisfies $\beta(1 + F_1(K^\star, \mu_A\omega_A + \mu_B\omega_B) - \delta) = 1$, and neither $\omega_A w + (1/\beta - 1)k_A$ nor $\omega_B w + (1/\beta - 1)k_B$ is negative (which ensures non-negative consumption for both agents). That is, given the unique level of capital consistent with steady state, any distribution of this capital will be a constant equilibrium where each individual just consumes the wage plus the interest on the capital. This case, including the associated transitional dynamics, is discussed in detail in Chatterjee (1994).[20] This model has no predictions for long-run wealth inequality, other than to perpetuate whatever inequality initially prevails. This result is robust to adding a proportional tax on capital income (with lump-sum rebates).

Heterogenity in critical places In contrast, assume that $\beta_A > \beta_B$. Then there is no steady state, but asymptotically there is extreme wealth inequality: agent A owns the entire capital stock plus a claim on agent B such that the latter has zero consumption. Intuitively, the relatively impatient agent B borrows early on and then pays back later. Now, the model has predictions, and they are stark. The same stark outcome would hold asymptotically if the two agents had the same discount factors but different returns on their capital: $r_A > r_B$; we can assume that this is achieved by means of a proportional tax on agent B's capital income and lump-sum transfers of the proceeds. Again, agent A would hold all the wealth asymptotically.

[19] The consideration of technological change gives slightly different results but does not materially affect the key discussion in what follows.

[20] Notice that $u_A(\cdot)$ need not equal $u_B(\cdot)$ for this result to hold.

Consider yet another case, where $\beta_A = \beta_B$ and $r_A = r_B$ but where there is a progressive tax rate on capital income. Assume first that this rate is strictly increasing in capital income. Then there is again a sharp prediction, but one with full equality: the only situation in which both agents' Euler equations can hold is that where they both have the same capital income and, therefore, the same levels of capital. A second case of interest obtains when the tax rate is weakly increasing in capital income, with flat sections. Then long-run inequality involves a unique total capital stock in steady state but a range of distributions of this stock—such that both agents remain within the same tax bracket.

Risk Relative to these results, let us consider stochastic earnings. First, consider the case where the total effective amount of labor is always constant but where all of the A agents receive the same shock and all of the B agents receive another shock; thus, by construction, there is perfect negative correlation between the shocks of the two agents. Under complete markets, i.e., when agents can fully insure, we obtain the same predictions for wealth inequality as above—in all the different subcases. In other words, random incomes do not matter per se.

However, when earnings are not fully insured, this result no longer holds. In particular, in the Bewley-Aiyagari-Huggett settings, there is only one asset and a constraint on borrowing and hence perfect consumption smoothing is not possible; there is, instead, "precautionary saving". Moreover, in all the cases discussed above—no heterogeneity, different discount factors, different returns, progressive income taxation—the model typically has a sharp long-run prediction: there is a unique, and non-degenerate, steady-state wealth distribution. Intuitively, given that future earnings are random and cannot be traded away unrestrictedly early on, relatively impatient consumers cannot end up in eternal poverty because their wage income will always bounce back, hence eliminating the extreme wealth inequality predicted under complete insurance/no earnings risk. Similar intuition applies in the other cases.

In the case with idiosyncratic, uninsurable risks, notice that partial-equilibrium analysis too becomes interesting. For example, a lowering of the risk-free interest rate at which agents save will have smooth effects on the average long-run wealth level held by a household, as well as on its ergodic distribution of wealth more generally. This contrasts with the "infinitely elastic" supply of household saving under complete markets/no earnings risk around the point where the interest rate equals the discount rate (where the long-run saving is zero (infinity) if the interest rate is lower (higher) than the discount rate by ever so little).

Comparative statics under idiosyncratic risk and incomplete markets A key purpose of the present subsection is to illustrate, with some examples, how the variance of earnings shocks can influence steady-state inequality in the incomplete-markets settings. In later sections, we will also comment on other types of comparative statics (e.g., with regard to the randomness in returns or in discount factors).

Suppose one departs from the case with a zero earnings variance and then increases it infinitesimally. How will steady-state wealth inequality then be affected? Under homogeneity in preferences and returns, long-run wealth inequality can go either up or down—depending on its starting position. If the starting position is the case with full equality, earnings volatility will necessarily increase wealth inequality in the long run, but if the starting position is at one of the extremes, wealth inequality will necessarily fall.

In the cases with either different discount factors or different person-specific returns, an increase in earnings volatility above zero must decrease wealth inequality in the long run. The result that more earnings risk can lower wealth inequality is perhaps not intuitive at first but with more risk one is further from the frictionless outcome, which is always extreme inequality in these cases.[21] Of course, higher earnings inequality can also increase long-run wealth inequality in these models, mechanically or because taxation is progressive (where absent shocks there is long-run equality). Kaymak & Poschke (2016) do report this finding and their framework is precisely one without return or discount-rate heterogeneity.

Non-dynastic households Finally, let us comment on how departures from dynastic models affect long-run inequality. The general answer is that it depends on what the bequest function looks like. If households derive utility from bequeathing, then if the associated function happens to look exactly like the value function in the associated dynastic household case—which would require it to also depend on any current idiosyncratic shock—then we have the same predictions as above, except insofar as we perform comparative statics.[22] If the bequest function, instead, is more or less curved than the associated value function, one would (heuristically) obtain less or more wealth inequality to be passed on from generation to generation; if the bequest function does not take the earnings state into account one would limit precautionary saving (to within one's own life). Absent definitive microeconomic estimates of bequest functions, we consider the dynastic structure a reasonable middle ground.

In the next sections, we describe our model economy. As advertised, the basic building block is the framework in Aiyagari (1994), on top of which we add several layers of complexity to account for the empirical evidence on earnings and return heterogeneity. The earnings process centers around a persistent and temporary component, augmented by a Pareto tail. The return on capital is stochastic. Both the mean and the dispersion of returns depend on the level of accumulated assets, a specification that can be interpreted as the reduced form of a full model of portfolio choice. Furthermore, the benchmark model also features stochastic discount rates. Let us now describe each component separately.

4.2 Consumers

Time is discrete and there is a continuum of infinitely lived, ex ante identical consumers (dynasties).[23] Preferences are defined over infinite streams of consumption with von Neumann-Morgenstern utility in constant relative risk aversion (CRRA) form:

$$u(c) = \frac{c^{1-\gamma}}{1-\gamma}. \tag{1}$$

In period t, a consumer discounts the future with an idiosyncratic stochastic factor β_t that is the realization of a Markov process characterized by the conditional distribution $\Gamma_\beta(\beta_{t+1}|\beta_t)$, giving rise to the following

[21] As an example, Krusell et al. (2009) shows that the removal of aggregate risk, which also involves a lowering of idiosyncratic risk, raises long-run wealth inequality quite significantly (since in that framework different households have different discount factors, so that the removal of idiosyncratic risks took us closer to the no-risk, extreme long-run inequality outcome).

[22] The bequest function not depending on the current idiosyncratic shock amounts to not letting bequests be influenced by the future income (shocks) faced by the offspring.

[23] To save on notation, we drop household subscripts from now on.

objective:

$$\max_{(c_t)_{t=0}^{\infty}} \left\{ u(c_0) + \mathbb{E}_0 \left[\sum_{t=1}^{\infty} \prod_{s=0}^{t-1} \beta_s u(c_t) \right] \right\}. \tag{2}$$

Labor supply is exogenous. Each period t, a consumer supplies a stochastic amount $l_t = l_t(p_t, \nu_t)$ of efficiency units of labor to the market that depends on a persistent component $p_t \sim \Gamma_p(p_t|p_{t-1})$ and a transitory component $\nu_t \sim \Gamma_\nu(\nu_t)$. Taking as given a competitive wage rate w_t, her earnings are $w_t l_t$.

Asset markets are incomplete, consumers cannot fully insure against idiosyncratic shocks. In the model, the only endogenous choice is the overall level of savings a_t. The gross return on it is

$$1 + \underline{r}_t + r_t^X(a_t) + \sigma^X(a_t)\eta_t, \tag{3}$$

where $\underline{r}_t$ is an aggregate return component, $r_t^X(\cdot)$ and $\sigma^X(\cdot)$ are functions that control mean and standard deviation of excess returns, and η_t is an i.i.d. standard normal idiosyncratic shock. The excess return schedule should be viewed as the reduced form of an implicit portfolio choice model, where the optimal choice is allowed to depend on the overall wealth level, albeit not on other persistent state variables. In addition to heterogeneity, this specification allows for a limited amount of return persistence: in the cross-section of all agents in this economy, returns are persistent because wealth is, but conditional on the level of wealth, returns are uncorrelated over time.[24]

The decision problem of the consumer can be stated in recursive form as follows:

$$V_t(x_t, p_t, \beta_t) = \max_{a_{t+1} \geq \underline{a}} \{ u(x_t - a_{t+1}) + \beta_t \mathbb{E}\left[V_{t+1}(x_{t+1}, p_{t+1}, \beta_{t+1}) | p_t, \beta_t \right] \} \tag{4}$$

$$\text{subject to } x_{t+1} = a_{t+1} + y_{t+1} - \tau_{t+1}(y_{t+1}) + (1 - \tilde{\tau}_{t+1})\tilde{y}_{t+1} + T_{t+1} \tag{5}$$

$$y_{t+1} = \left(\underline{r}_{t+1} + r_{t+1}^X(a_{t+1}) \right) a_{t+1} + w_{t+1} l_{t+1}(p_{t+1}, \nu_{t+1}) \tag{6}$$

$$\tilde{y}_{t+1} = \sigma^X(a_{t+1}) \eta_{t+1} a_{t+1} \tag{7}$$

Given cash-on-hand x_t (all resources available in period t), the optimal savings decision and the resulting value function depend solely on the persistent component of the earnings process p_t and the current discount factor β_t. Conditional on (p_t, β_t), the expectation is taken over (p_{t+1}, β_{t+1}) as well as the transitory shocks to earnings ν_{t+1} and the return on capital η_{t+1}. Ordinary gross income y_t is subject to a non-linear income tax $\tau_t(\cdot)$, while there is a flat (capital gains) tax $\tilde{\tau}_t$ on the mean-zero idiosyncratic return component.[25] Each consumer receives a uniform lump-sum transfer T_t.

[24] Fagereng et al. (2020) and Bach et al. (2019) find not only heterogeneity but persistence in idiosyncratic asset returns. However, especially Bach et al. (2019) find that a good portion of this persistence stems from richer consumers bearing more aggregate risk, which we do not model here. Furthermore, we find below that we can replicate the wealth distribution in 1967, even in its remotest tails, quite accurately without genuine persistence in idiosyncratic returns.

[25] In the presence of a progressive income tax, sophisticated agents would seek to smooth capital income over time. For tractability reasons, instead we impose a flat tax on the mean-zero stochastic capital income component.

4.3 Production, government, and equilibrium

Firms are perfectly competitive and can be described by an aggregate constant returns to scale production function $F(K_t, L)$ that yields a wage rate per efficiency unit of labor $w_t = \frac{\partial F(K_t,L)}{\partial L}$ as well as an (average) market return on capital $r_t = \frac{\partial F(K_t,L)}{\partial K} - \delta$, where $\delta \in (0,1)$ is the depreciation rate. Aggregate labor supply L is normalized to one throughout.

As in Aiyagari (1994), aggregate capital K_t equals the average of consumers' asset holdings a_t in equilibrium. Thus, the production side is rather standard, and aggregate capital income, net of depreciation, is $r_t K_t$. However, in case there is a non-trivial excess return schedule $r_t^X(\cdot)$, individual capital income is not proportional to asset holdings (i.e., not even the expectation of it). Thus, in order for capital market clearing, a second condition has to hold, namely that aggregate capital income equals the average over individual capital income. Both $r_t^X(\cdot)$ and $\sigma^X(\cdot)$ are treated as exogenous objects (that will be taken from the data), thus the scalar $\underline{r}_t$ is the second aggregate equilibrium object, beside K_t. Note that $\underline{r}_t$ is not solely a function of K_t, but depends on the asset distribution as well.

The government redistributes aggregate income by means of a uniform lump-sum payment, which amounts to a constant fraction $\lambda \in [0,1]$ of aggregate tax revenues. The remainder is spent in a way such that marginal utilities of agents are not affected. Since revenues from the flat capital gains tax net out to zero in the aggregate, we omit them from the government budget constraint for simplicity.

Given time-invariant excess return schedules $r^X(\cdot)$ and $\sigma^X(\cdot)$, a steady-state equilibrium of this economy is characterized by a market clearing level of capital $K^\star$, an aggregate return component $\underline{r}^\star$, and a lump-sum transfer $T^\star$ such that:

(i) factor prices are given by their respective marginal products $w^\star = \frac{\partial F(K^\star,1)}{\partial L}$ and $r^\star = \frac{\partial F(K^\star,1)}{\partial K} - \delta$;

(ii) given $\underline{r}^\star$, $w^\star$, and $T^\star$, consumers solve the stationary version of their decision problem, giving rise to an invariant distribution $\Gamma(a, p, \beta, \nu, \eta)$;

(iii) the government redistributes a fraction λ of total tax revenues, i.e.,

$$T^\star = \lambda \int \tau(\left(\underline{r}^\star + r^X(a)\right) a + w^\star l(p,\nu)) d\Gamma(a, p, \beta, \nu, \eta);$$

(iv) and capital markets clear, i.e.,

$$K^\star = \int a d\Gamma(a, p, \beta, \nu, \eta), \qquad \text{and}$$
$$r^\star K^\star = \int \left(\underline{r}^\star + r^X(a) + \sigma^X(a)\eta\right) a d\Gamma(a, p, \beta, \nu, \eta).$$

In the benchmark *perfect-foresight* transition experiment, we start the economy in period t_0 in some initial steady state, described by a parameter vector $\theta^\star$ and by the equilibrium objects $(K^\star, \underline{r}^\star, T^\star)$. The vector $\theta^\star$ parametrizes the tax schedule, the excess return schedule, and the earnings process. Agents are fully surprised and learn about a new exogenous environment $(\theta_t)_{t=t_0+1}^{t_1}$ that will prevail over some transition period $t = t_0 + 1, t_0 + 2, ..., t_1$. From t_1 onwards, the exogenous environment will once again be constant and equal to θ_{t_1}. In a perfect-foresight equilibrium, agents are fully informed about future equilibrium objects $(K_t, \underline{r}_t, T_t)_{t=t_0+1}^{\infty}$ too and optimize accordingly. Capital markets clear and the fraction of tax revenues λ that is redistributed is fixed.

In an alternative *myopic* transition experiment, agents are surprised about the new exogenous environment and equilibrium prices every period. That is, in period $t = t_0, t_0+1, ..., t_1-1$, given a distribution $\Gamma_t(x_t, p_t, \beta_t)$, they choose a savings decision rule, $a_{t+1} = g_t(x_t, p_t, \beta_t)$, assuming that both θ_t and $(\underline{r}_t, w_t, T_t)$ will prevail forever. In period $t+1$, they are accordingly surprised that: one, the exogenous environment has changed to θ_{t+1}; and, two, that equilibrium factor returns $(\underline{r}_{t+1}, w_{t+1})$ and transfers T_{t+1} result from capital-market clearing and government-budget balance in period $t+1$.[26] These two informational structures are, of course, extreme. We chose them because we expect them to bracket a range of informational assumptions. Given that the results, as will be reported below, turn out to be very similar across the two structures, we are confident that our findings are robust to other variations in this dimension.

5 The right tail of the wealth distribution: approximately Pareto

In this section, we briefly explain the main mechanism that leads to a "fat" Pareto-shaped right tail in the wealth distribution. The same mechanism is at play in the much simpler stochastic-β model originally proposed in Krusell & Smith (1998).

Formally, we make use of a mathematical result on random growth by Kesten (1973): consider a stochastic process

$$a_t = s_t a_{t-1} + \epsilon_t, \tag{8}$$

where s_t and ϵ_t are (for our purposes positive) i.i.d. random variables. If there exists some $\zeta > 0$ such that $\mathbb{E}[s^\zeta] = 1$ as well as $\mathbb{E}[\epsilon^\zeta] < \infty$, then a_t converges in probability to a random variable A that satisfies $\lim_{a\to\infty} Prob(A > a) \propto a^{-\zeta}$, i.e., the right tail of the stationary distribution has a Pareto shape.[27]

In a setup like ours, it turns out—as we discuss in some more detail below—that s is the asymptotic marginal propensity to save out of initial-period asset holdings. Moreover, this propensity is random, whence it obtains a time subscript. In a basic model with only discount-factor randomness, s varies precisely with β; this turns out to be a property already of the model in Krusell & Smith (1998) designed

[26] That is, (r_{t+1}, w_{t+1}) are the marginal products of the net production function $F(K_{t+1}, 1) - \delta K_{t+1}$, where

$$K_{t+1} = \int g_t(x_t, p_t, \beta_t) d\Gamma_t(a_t, p_t, \beta_t, \nu_t, \eta_t),$$

$\underline{r}_{t+1}$ is given by

$$\underline{r}_{t+1} = r_{t+1} - \frac{1}{K_{t+1}} \int \left(r_{t+1}^X(a_{t+1}) + \sigma^X(a_{t+1})\eta_{t+1} \right) a_{t+1} d\Gamma_{t+1}(a_{t+1}, p_{t+1}, \beta_{t+1}, \nu_{t+1}, \eta_{t+1}),$$

and

$$T_{t+1} = \lambda \int \tau_{t+1}\left(\left(\underline{r}_{t+1} + r_{t+1}^X(a_{t+1}) \right) a_{t+1} + w_{t+1} l_{t+1}(p_{t+1}, \nu_{t+1}) \right) d\Gamma_{t+1}(a_{t+1}, p_{t+1}, \beta_{t+1}, \nu_{t+1}, \eta_{t+1}),$$

where Γ_{t+1} is the distribution in period $t+1$ generated by the period-t distribution Γ_t and the decision rule g_t.

[27] The exact conditions as well as a very accessible treatment can be found in Gabaix (2009).

to match the wealth distribution, though the β-distribution there is quite stripped down. In the present somewhat augmented model, s_t also varies with the idiosyncratic return to wealth, η_t. Random earnings appear in the linear approximation through the error term ϵ_t. Crucially, in this class of models, optimal saving decisions are asymptotically, with increasing wealth, linear in economies with idiosyncratic risk and incomplete markets.[28]

Assuming a fixed discount rate, Carroll & Kimball (1996) prove in a finite-horizon setting that the consumption function is concave under hyperbolic absolute risk aversion, which comprises most commonly used utility functions (e.g., CRRA). Hence, the savings rule is convex. However, as household wealth increases, the convexity in the savings rule becomes weaker and weaker.[29] Intuitively, as wealth grows large consumers can smooth consumption more and more effectively. Moreover, with CRRA preferences decisions rules are exactly linear in the absence of risk (or with complete markets against such risk). The slope is then larger (smaller) than one as the discount rate is smaller (larger) than the interest rate. In the recent literature on the Pareto tail in the wealth distribution, either saving rates or returns to capital (or both, as in this paper) are assumed to vary randomly across consumers. Saving rules are then asymptotically linear with random coefficients: Benhabib et al. (2015) show analytically that in this case the unique ergodic wealth distribution has a Pareto distribution in its right tail.

Figure 2 shows the marginal propensity to save out of capital holdings (denoted k in the figure) arising from the stochastic-β model under study in the present paper.[30] As discussed above, the marginal propensity to save increases in wealth, holding earnings constant, and asymptotes to a constant that depends on the consumer's discount factor. Figure 3 displays the tail behavior of the stationary wealth distribution. In line with the theoretical results in Benhabib et al. (2015), the logarithm of its counter-cumulative distribution function becomes linear in the logarithm of assets as assets grow large, indicating that the right tail of the distribution follows a Pareto distribution.

In light of this result, it is worth noting that the model in Castañeda et al. (2003)—which generates substantial wealth inequality using an earnings process featuring a low-probability but transient very-high-earnings state—does not deliver a Pareto tail in wealth. In this model, in which consumers have a common discount rate, marginal propensities to save do not vary but instead converge to the same constant, independently of the level of earnings and as a result the steady-state distribution of wealth does not feature a Pareto tail. This model can deliver such a Pareto tail, however, if the earning process itself has a Pareto tail. In the absence of randomness in either discount rates or returns, however, the wealth distribution inherits not only the Pareto tail of the earnings distribution but also its Pareto coefficient. Because earnings are considerably less concentrated than wealth, the resulting tail in wealth is too thin to match the data in such an alternative model.

[28] In fact, the decision rules are almost linear for all but the very poorest agents, i.e., those close to the borrowing constraint. For this reason, approximate aggregation as introduced in Krusell & Smith (1998) typically works very well.

[29] A direct proof for a two-period problem can be found in Krusell & Smith (2006); Carroll (2012) proves the asymptotic linearity of the savings rule in a finite-horizon problem as the horizon grows large.

[30] The graphs in this section are derived from a simplified model with a flat tax, to focus on the main mechanism.

Figure 2: Asymptotic marginal propensity to save

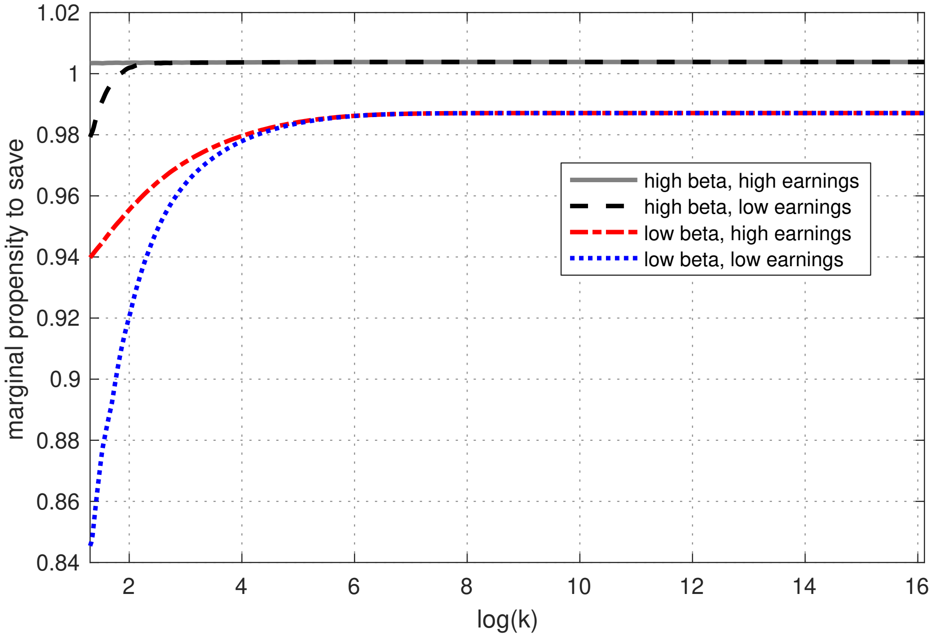
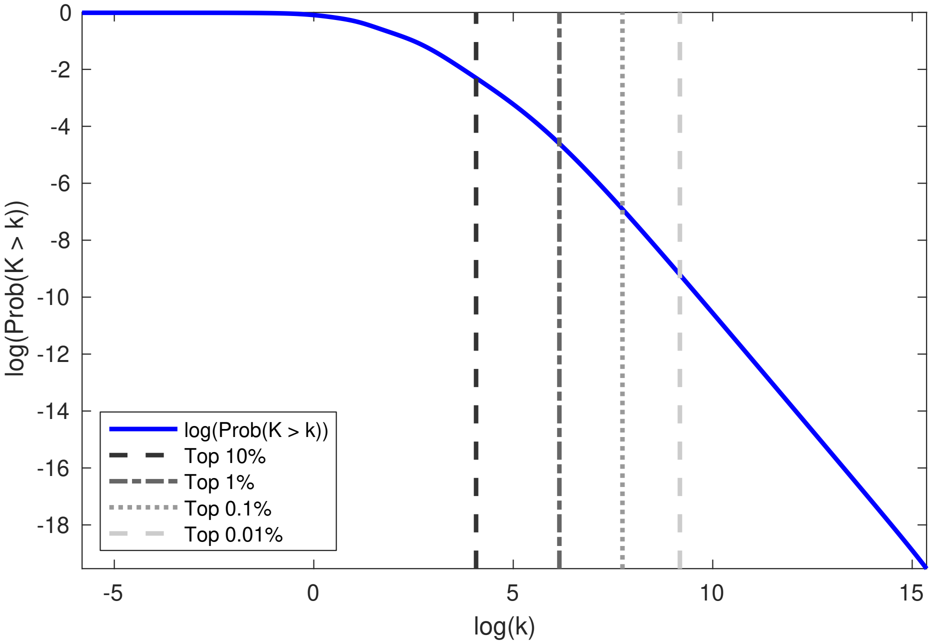


Figure 3: Pareto tail of the wealth distribution



6 Calibration

In this section, we describe how we calibrate our model economy. As indicated in Figure 1, the U.S. wealth distribution was roughly stable in the 1950s and 1960s, as was tax progressivity. This, together with the fact that some of our time series estimates start in 1967, make this year a natural initial steady state. We set the model period to a year to conform to the tax system. Overall, the strategy is to use observables to select the structural model parameters to the largest extent possible; the key observables are the earnings process, the tax system, and the households' portfolio and return structures. To the extent not all the wealth inequality can be accounted for this way, we then calibrate the discount factor process to match the 1967 wealth distribution as completely as possible (given the parsimonious process for discount factors and the multidimensionality of the wealth distribution, a full match is of course not feasible). As we shall see, we present two main cases, in one of which there is no discount-factor heterogeneity at all (and the main results differ only marginally between these two cases).

6.1 Basic parameters

We parameterize the production technology and utility function using standard functional forms and parameters. The (gross) production function is given by $F(K, L) = K^{\alpha}L^{1-\alpha}$. The capital share is set to $\alpha = 0.36$ and depreciation to $\delta = 0.048$ annually. In an extension (see Section 9.1), we check the sensitivity of our results to using a constant-elasticity-of-substitution production function with (gross) elasticity greater than one. The coefficient of relative risk aversion, γ, is set to 1.5.

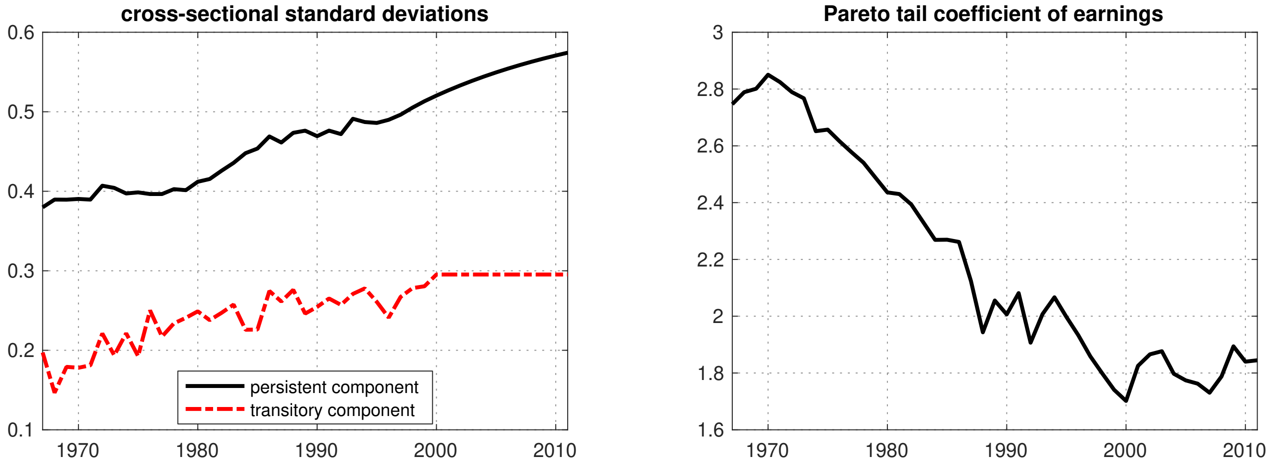
6.2 The earnings process

The earnings process is based on the traditional log-normal framework with $l_t(p_t, \nu_t) = \exp(p_t + \nu_t)$. That is, we assume that the persistent component p_t of the earnings process follows a Gaussian AR(1) process with parameters (ρ^P, σ_t^P). The autocorrelation coefficient, ρ^P, is fixed over time, while the innovation standard deviation varies. Likewise, the transitory component ν_t is also assumed to be normally distributed with standard deviation σ_t^T. We use estimates by Heathcote et al. (2010) that span the period 1967–2000 and assume that the time-varying variances of the innovations are constant thereafter. The left panel of Figure 4 displays the resulting cross-sectional dispersion. The estimates show a significant increase in earnings risk for both components.

As is well known, the resulting log-normal cross-sectional distribution of earnings understates the concentration of top labor income quite severely. Because the observed increase in top labor income shares is potentially an important explanation for the observed increase in wealth inequality at the top, we augment the framework for the top 10% earners in such a way that we can directly match the fraction of labor income going to the top 10%, top 1%, top 0.1% and top 0.01%. In concrete terms, we posit $l_t(p_t, \nu_t) = \psi_t(p_t)\exp(\nu_t)$, where

$$\psi_t(p_t) = \begin{cases} \exp(p_t) & \text{if } F_{p_t}(p_t) \leq 0.9, \\ F^{-1}_{Pareto(\kappa_t)}\left(\frac{F_{p_t}(p_t)-0.9}{1-0.9}\right) & \text{if } F_{p_t}(p_t) > 0.9. \end{cases} \tag{9}$$

Figure 4: Earnings process ingredients



Left panel: cross-sectional standard deviation of temporary and persistent component of labor income from Heathcote et al. (2010). Right panel: Pareto tail coefficient for top 10% of labor income distribution; calibrated to replicate updated Piketty & Saez (2003) series.

$F_{p_t}(\cdot)$ is the cdf of p_t and $F^{-1}_{Pareto(\kappa_t)}(\cdot)$ the inverse cdf for a Pareto distribution with lower bound $F^{-1}_{p_t}(0.9)$ and shape coefficient κ_t. Effectively, we thus assume that top earnings are spread out according to a (scaled) Pareto distribution, while earnings for the majority of workers are distributed according to a log-normal distribution. The Pareto tail coefficient on labor income κ_t is then one additional free parameter to calibrate in each year. We use estimates on top wage shares from an updated series by Piketty & Saez (2003) spanning 1967–2011 as calibration targets. The right panel of Figure 4 displays the calibrated Pareto tail coefficient κ_t and Figure 5 displays the resulting top labor income shares. That we can match top labor income shares very well using just a single parameter in each year (i.e., the tail coefficient) simply reflects the fact that the Pareto distribution is a very good description of the cross-sectional earnings distribution at the top.

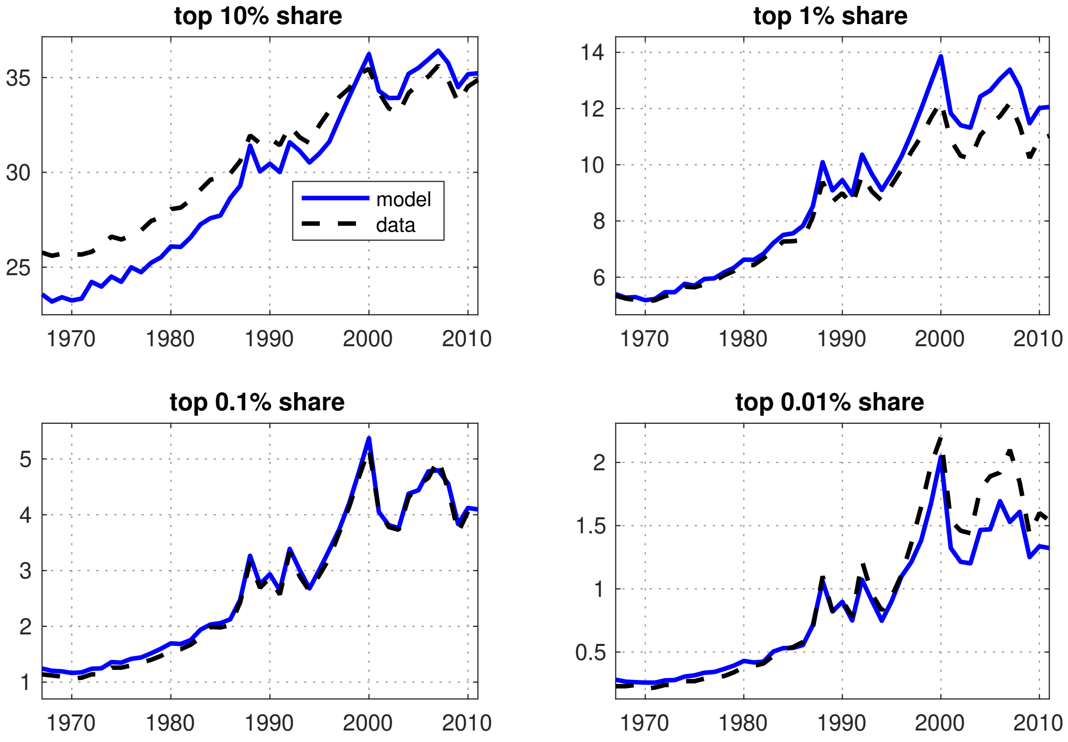
We do not explicitly model unemployment, nor voluntary non-employment or retirement. We do, however, introduce a zero-earnings state, occurring with probability $\chi = 0.075$ independently of (p_t, ν_t) and over time, reflecting both long-term unemployment and shocks that trigger temporary exit from the labor force. This probability is calibrated, together with a borrowing constraint amounting to roughly one yearly lump-sum transfer, so that the initial steady-state wealth distribution matches both the share of wealth held by the bottom 50% and the fraction of the population with negative net wealth.

6.3 Tax system

The progressivity of the U.S. tax system has decreased substantially over the model period. To account for these changes, we use estimates on federal effective tax rates by Piketty & Saez (2007) for the period 1967–2000, keeping them constant thereafter. These comprise the four major federal taxes: individual income, corporate income, estate and gift, and payroll taxes.[31] Piketty & Saez (2007) calculate effective

[31] Given that our model abstracts from the life cycle, it is appropriate to include the estate tax in the tax on total income, thus effectively smoothing out the incidence of this tax over the life cycle. Ignoring the estate tax would mean omitting a

Figure 5: Top labor income shares in %



Data source: updated Piketty & Saez (2003) series.

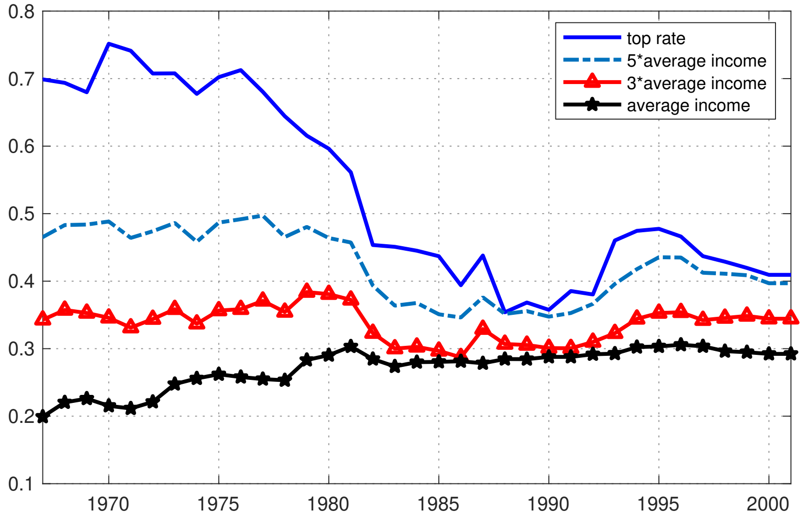
average tax rates for eleven income brackets, with a particularly detailed decomposition for top income groups (up to the top 0.01%). We translate this data to our model by means of a step-wise tax function $\tau_t(\cdot)$ with eleven steps. For each bracket, the threshold is set to match its income share in the data and the marginal tax rate such that the resulting average tax rate aligns with the data. Figure 6 shows that the U.S. tax system has indeed become much less progressive over the model period.

In our model, taxes $\tau_t(y_t)$ are a function of total ordinary income y_t, defined as the sum of labor income and the deterministic part of capital income. A weakness of our calibration is that we do not have separate tax rates for different sources of income, but a strength is that we use effective tax rates, thereby accounting for tax avoidance and changing portfolio composition to the extent that these vary systematically with income.

The stochastic part of capital income is uncorrelated over time, and equals zero in expectation for every agent. Especially at the top end of the (capital) income distribution, with sizable return risk and thus sizable year-to-year capital income fluctuations, agents have strong incentives to smooth reported capital income over time if the tax function is progressive. To avoid dealing with this issue in full detail, we use a time-varying flat tax $\tilde{\tau}_t$ for this part of capital income. In particular, we use an annual time

major source of decreasing tax progressivity. Piketty & Saez (2007) assume further that the corporate income tax burden falls entirely (and uniformly) on capital income. They argue that this is a middle-ground assumption (regarding the resulting tax progressivity) between assuming that the tax falls solely on shareholders at one extreme and assuming that it is effectively born by labor income at the other extreme.

Figure 6: Imputed marginal tax rates for selected total income levels



Data source: Piketty & Saez (2007).

series on the average effective capital gains tax.[32]

To account for government transfers, we introduce a social safety net in the simplest possible way by assuming that each agent receives an (untaxed) lump-sum transfer T_t every period, its size being a constant fraction $\lambda = 0.6$ of tax revenues.[33] The presence of a lump-sum transfer implies that the resulting net tax and transfer system is substantially more progressive than the gross tax system, as we show in Appendix C.

Note that the income tax does not distort labor supply in our setting, since we assume the latter is exogenous. This simplification is obviously not a good one for understanding the welfare consequences of changes in tax rates, but because our current focus is on wealth accumulation and its distribution in the population we do not think that it is a major shortcoming.

6.4 Idiosyncratic returns to capital

The idiosyncratic return component depends on the overall wealth level a_t. In recent work, both Bach et al. (2019), using Swedish administrative data, and Fagereng et al. (2020), using Norwegian administrative data, document a strong relation between a household's overall wealth and return characteristics. These papers disagree somewhat in their conclusions as to whether differential returns can be fully explained by differential portfolio choice. The possibility that different households have different skills in finding returns (an interpretation made in Fagereng et al. (2020)) is particularly radical relative to the traditional

[32] The time series is published in U.S. Department of the Treasury (2016). This is a slight approximation to the actual, historical, U.S. tax schedule for capital gains, which features rates that vary across asset categories, amount of time the asset was held, and also overall income. The capital gains tax schedule has been slightly progressive as well, though much less so than the one on ordinary income.

[33] About 60% of total federal outlays are mandatory spending, the bulk of it on Social Security, Medicare, Medicaid, and income security programs (CBO, 2015). The remainder is spent on the Department of Defense and other government agencies as well as on interest payments.

finance literature. Although we do not want to rule out this hypothesis is true, our calibration strategy is to follow the work of Bach et al. (2019). In particular, we calibrate the schedules of mean excess returns $r_t^X(a_t)$ and return dispersions $\sigma^X(a_t)$ such that they represent an approximation to the reduced form of an underlying portfolio choice model.[34]

The mean excess return schedule is computed as

$$r_t^X(a_t) = \sum_{c \in C} w_c(a_t) \left(\bar{r}_{c,t} + \tilde{r}_c^X(a_t) \right), \tag{10}$$

where $w_c(a_t)$ is the portfolio weight on asset class c, $\bar{r}_{c,t}$ is the aggregate excess return on asset class c, and $\tilde{r}_c^X(a_t)$ an idiosyncratic component that accounts for within-asset class return heterogeneity. We consider four asset classes: a risk-free asset, public equity, private equity, and housing. The schedules for portfolio weights $w_c(\cdot)$ and within-asset class heterogeneity $\tilde{r}_c^X(\cdot)$ are fixed over time. We base them on data from Bach et al. (2019), who report a detailed breakdown up to the top 0.01%.[35] Aggregate excess returns $\bar{r}_{c,t}$ are time-varying and based on aggregate U.S. data. In particular, for public and private equity, we use estimates from Kartashova (2014), who documents a premium for private equity over public equity. For housing, we model the financial return as the sum of capital gains and imputed rent. For the capital gains term, we rely on the national Case-Shiller home price index.[36] In the initial and eventual steady states, we assume that house prices grow at the rate of overall inflation, in line with long-run evidence. We assume that the imputed rent term is fixed over time; we set it to 5.33%, the U.S. time average reported in Jordà, Knoll, Kuvshinov, Schularick, & Taylor (2019). Note that the level of the excess return schedule $r_t^X(\cdot)$ is irrelevant, as the endogenous aggregate return component $\underline{r}_t$ adjusts for market clearing. In other words, only differences in returns across asset classes, and within, are treated as exogenous.

The schedule of idiosyncratic return dispersion is computed as

$$\left(\sigma^X(a_t)\right)^2 = \sum_{c \in C} \left(w_c(a_t) \tilde{\sigma}_c^X(a_t) \right)^2, \tag{11}$$

where the idiosyncratic standard deviation of the return on asset class c, $\tilde{\sigma}_c^X(\cdot)$, is fixed over time but allowed to depend on the wealth level. For private and public equity, we again rely on Bach et al. (2019). For housing, we set the standard deviation to 0.14 across the wealth distribution, based on the observed volatility of individual house prices in the U.S.[37]

Figure 7 summarizes the excess return schedule in the 1967 steady state. Full details are relegated to the appendix (see Table 6). As we explain below, using the unadjusted schedules results in too much wealth inequality at the very top. For this reason, in the benchmark model we scale down the standard deviation of private equity across the board by a factor $\phi = 0.52$. As can be seen in Figure

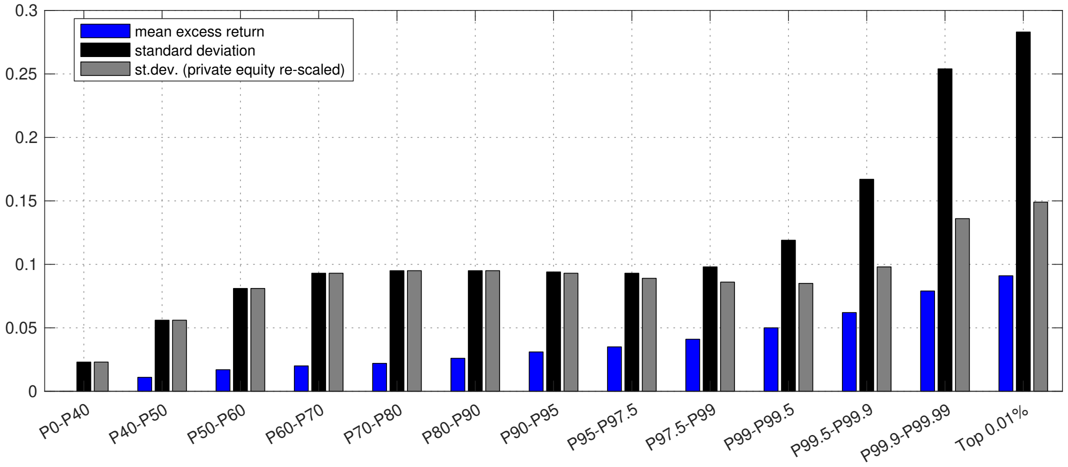
[34] In a fully rational portfolio choice model, optimal allocations would depend not only on the wealth level, but also on our other persistent states (in particular, the persistent component of earnings). The data we use does not allow for this level of detail.

[35] In Appendix E, we argue that U.S. data is comparable, to the extent that it is available. We also demonstrate that relative portfolio differences across wealth groups are quite stable over time.

[36] The series is accessible at `http://www.econ.yale.edu/~shiller/data/Fig3-1.xls`.

[37] See Piazzesi & Schneider (2016).

Figure 7: Schedule of excess returns

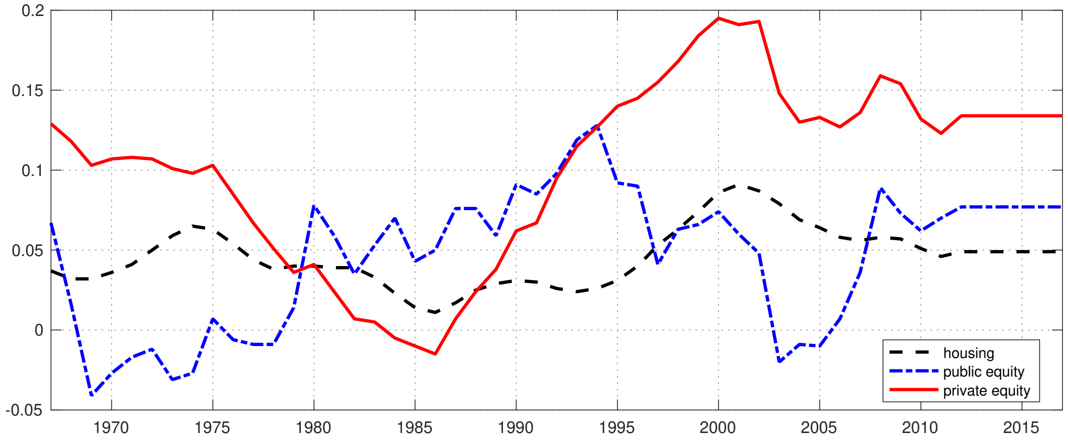


Source: own computations based on Bach et al. (2019), Kartashova (2014), Jordà et al. (2019), Piazzesi & Schneider (2016), Case-Shiller home price index.

7, this adjustment reduces in particular the volatility of the portfolio standard deviation for the top 1% of the wealth distribution, and consequently reduces the thickness of the extreme right tail to a level commensurate with data.

Over time, only aggregate returns by asset class $\bar{r}_{c,t}$ are varying. We use ten-year moving averages of realized aggregate returns for the transition, displayed in Figure 8. These are expressed relative to the return on the base category, the risk-free asset.

Figure 8: Aggregate excess returns



Source: own computations based on Kartashova (2014), Case-Shiller home price index. The series correspond to 10-year moving averages of realized excess returns over the federal funds rate. The initial steady state in 1967 is based on time averages over 1960–1967. From 2012 onwards (including the new steady state), we use time averages over 2007–2017.

6.5 Idiosyncratic discount rates

We provide results for two model versions. In the first one, we do not rescale the standard deviation of private equity ($\phi = 1$) and we do not allow for preference heterogeneity. We refer to this as the single-β model. Thus, the only two free parameters to calibrate are then the borrowing constraint and the probability of the zero earnings state, which mostly affect the bottom end of the wealth distribution. Table 1 shows that, quite remarkably, the resulting invariant wealth distribution matches the data in 1967 quite well.[38]

While the single-β model reproduces the overall amount of wealth inequality, it overstates wealth concentration at the top end. To the extent that the wealth distribution has a Pareto tail to the right, this coefficient is pinned down by the ratio of the top 0.01% share to the top 0.1% share, or the ratio of the top 0.1% share to the top 1% share, both of which are roughly one-third in the data. In the single-β model, this ratio is increasing the further one moves out in the right tail, and stabilizing at a value that is by far too high. These findings motivate the specification of the benchmark model: a model that in addition allows for discount factor (β) heterogeneity and rescales the standard deviation of private equity returns by a factor ϕ. Intuitively, the discount-factor distribution affects the entire asset distribution, including the Pareto tail coefficient. More heterogeneity creates more wealth inequality. The standard deviation of private equity returns, on the other hand, mostly affects the very right tail, and thus the tail coefficient.

We use an AR(1) structure for the discount factor. Thus, from the perspective of dispersion in the benchmark model we have three parameters to calibrate: the variance and persistence of β and the scaling factor ϕ. First, we select the persistence of the β process based on what seems a priori reasonable given a generational structure. Second, we target two wealth-distribution statistics to obtain the remaining two variance elements (σ^β and ϕ): the Pareto tail coefficient and the fraction of total wealth held by the 10% richest. This identifies our parameters. We now describe the details.

We posit that β follows a Gaussian AR(1) process:

$$\beta_t = \rho^\beta \beta_{t-1} + (1-\rho^\beta)\mu^\beta + \sigma^\beta \epsilon_t^\beta, \qquad \epsilon_t^\beta \sim N(0,1).$$

Importantly, all these parameters are fixed over time (by varying them freely we could of course track the evolution of the wealth distribution at will). The mean discount factor determines the equilibrium capital-output ratio and we set it to $\mu^\beta = 0.944$ to match a ratio of capital to net output of about 4 in the initial steady state. The calibrated stochastic-β parameters are $\rho^\beta = 0.992$ and $\sigma^\beta = 0.0006$, implying that the standard deviation of the cross-sectional distribution of discount factors, which does not vary over time, is 0.0050. Moreover, the choice of ρ^β implies that roughly one third of the gap between a given discount factor and the average discount factor is closed within a generation.

To summarize the calibration of the benchmark model, Table 1 lists the values of the five parameters (persistence and standard deviation of the discount rates; standard deviation of return shocks; the

[38] The data on top wealth shares in Table 1 is from Saez & Zucman (2016), who use a capitalization method to calculate them. Because this method is unreliable for a breakdown of the bottom 90%, the other data moments are based on survey data (SCF and precursors); see Kennickell (2011).

Table 1: Matching the 1967 wealth distribution as a steady state

Parameter	ρ^β	σ^β	ϕ		$\underline{a}$	χ
Single-β Model	n.a.	(0.0)	(1.0)		-0.26	7.5%
Benchmark Model	0.992	0.0006	0.52		-0.22	7.5%
Target	Top 10%	Top 1%	Top 0.1%	Top 0.01%	Bottom 50%	Fraction $a < 0$
Data	70.8%	27.8%	9.4%	3.1%	4.0%	8.0%
Single-β Model	66.6%	23.7%	11.2%	7.2%	3.5%	7.3%
Benchmark Model	73.8%	27.4%	8.4%	3.2%	3.0%	6.6%

Data sources: Top wealth shares from Saez & Zucman (2016); bottom 50% and fraction of population with negative net wealth from Kennickell (2011) based on SCF.

borrowing constraint; and the probability of zero income) calibrated to match six features of the initial steady-state wealth distribution as closely as possible: the shares held by the top 10%, the top 1%, the top 0.1%, the top 0.01%, and the bottom 50% as well as the fraction of the population with negative net wealth. The fit is excellent at both ends of the distribution. Compared to the single-β model, the benchmark model matches the Pareto tail coefficient in addition.

Two comments are in order. First, when solving the model numerically we truncate the β and η distributions to ensure that the consumer's optimization problem is well-defined (with finite present-value utility) and that a stationary distribution of wealth emerges. Unlike in a standard Aiyagari economy without heterogeneity in preferences or returns, in our model some agents temporarily have discount rates that are smaller than the rate of return, a necessary condition for generating a Pareto tail in the wealth distribution based on discount-rate or return heterogeneity alone (see the discussion in Section 5). It follows that the support of the stationary wealth distribution is not bounded from above. In practice, we use a large enough upper bound in our numerical implementation so that the resulting truncation error is negligible.[39]

Second, if our goal were solely to match the Pareto coefficient in the right tail of the wealth distribution, it would be excessive to calibrate as many as five parameters to match features of the wealth distribution. But the tail coefficient is not a sufficient statistic for wealth inequality unless the entire distribution is (counterfactually) Pareto-shaped: even if, say, the top 1% of the wealth distribution can be described exactly by a Pareto distribution, the tail coefficient determines only the distribution of wealth within these top 1% but not the fraction of total wealth held by the top 1%.

7 Results I: steady-state wealth inequality

A first and, we believe, important contribution of the present paper is its comprehensive breakdown of long-run wealth inequality, on which Section 7.2 reports. Such a breakdown is also useful because it hints at what to expect from movements over time in some of the drivers of long-run wealth inequality—the

[39] Appendix A describes in detail our numerical procedure.

subject of Section 8. In Section 7.1, we first briefly relate to the relevant literature. Both these sections draw on, and to some extent reiterate, the earlier discussions in Sections 2 and 4.1.

7.1 Relations to the literature on long-run wealth inequality

In the basic Aiyagari (1994) setting, where steady-state earnings inequality is calibrated as an AR(1) process to PSID data, very little wealth inequality is generated. Intuitively, the very highest earners are well-insured and the interest rate is not high enough to maintain their asset levels: it is below the discount rate, because it is depressed by the precautionary saving of less well-insured households, and so they decumulate. Since then, the literature has thus had the challenge to come up with mechanisms that generate a greater accumulation, or maintenance, of wealth by the richest. Krusell & Smith (1998) propose discount-rate heterogeneity, so that the richest are rich because they choose to save at higher rates than others. Castañeda et al. (2003) propose a different earnings process, whereby there are extreme right-tail outcomes at the same time as the risk of very large drops in earnings for the extreme earners is non-negligible. Hence, precautionary saving operates in the right tail as well. Quadrini (2000) and Cagetti & De Nardi (2006) look at entrepreneurs (and occupational choice) specifically as a candidate richest group and argue that the returns to saving can be higher for high wealth levels. Relatedly, Campanale (2007) uses a return schedule that is simply increasing in wealth, motivated by the fact that wealthier households hold more stock. Giving the bequest function a low curvature can also help (see, e.g., Cagetti & De Nardi (2009)). More recently, idiosyncratic return heterogeneity has been explored by a number of papers, as discussed in Section 2.

We clearly view discount-factor heterogeneity as realistic but, since we do not have fully reliable measurements of it, it plays only a residual role. We view the Castañeda et al. (2003) approach as interesting but somewhat problematic because it does not rely on independent, direct measurement of the earnings process—some features are selected to match wealth inequality—and the implied right-tail features of the earnings distribution are, in fact, too extreme compared to data. Moreover, as already discussed, it is not consistent with a Pareto-shaped right tail in wealth. Our Table 1 above shows that, quite encouragingly, given the observables we use, it is no longer necessary to resort to residual explanations (such as preference heterogeneity or a non-altruistic bequest function) to generate realistic wealth inequality.[40] In fact, the right tail of the wealth distribution is too thick relative to the data, so that, as explained in Section 6.5, we rescale the observed cross-sectional variation in returns to private equity and instead introduce a small amount of heterogeneity in discount factors to provide an even better fit to the entire distribution of wealth.

The next section will detail how each of the factors behind long-run wealth dispersion matter, but let us already emphasize that portfolio and return heterogeneity is key. What our present paper does not provide is a deeper theory either of portfolio choice or of return differences across assets. The latter have plagued the macro-finance literature since Mehra & Prescott (1985), but to understand the former is at least as challenging. One should therefore view our encouraging results here as far from satisfactory;

[40] One can view a departure from the dynastic structure, by "freeing up" the bequest function, as an alternative very similar to discount-factor heterogeneity.

Table 2: Contribution of various channels for steady state wealth inequality in the benchmark model

#		top 10%	top 1%	top 0.1%	top 0.01%	Gini
1	β-heterogeneity	8.8%	7.7%	3.8%	2.0%	0.050
2	earnings heterogeneity	−27.5%	−17.8%	−9.5%	−6.4%	−0.173
3	persistent	−5.0%	−7.5%	−4.2%	−2.9%	0.009
4	transitory	−11.6%	−4.3%	−1.7%	−0.9%	−0.109
5	tax progressivity	−21.3%	−61.8%	−71.2%	−67.1%	−0.148
6	return heterogeneity	29.5%	18.4%	6.6%	2.8%	0.192
7	mean differences	25.8%	16.7%	6.0%	2.6%	0.174
8	return risk	0.7%	2.2%	3.3%	2.5%	0.004

rather, we now need to turn to household finance and investments as a key area for understanding long-run wealth inequality.

7.2 Decomposing wealth inequality in the benchmark model

How much does each of the various sources of heterogeneity contribute to wealth inequality in the benchmark economy? To answer this question, we start from the benchmark model, shut down one channel at a time, and report on the general equilibrium differences in Table 2. These counterfactual exercises also give clues as to how the dynamics will work out—but of course will not help us understand the speed of these dynamics.

The first row in the table corresponds to a counterfactual in which discount factor heterogeneity is removed ($\sigma^{\beta} = 0$). Then, e.g., the top 10% wealth share decreases from 73.8% (the value in the benchmark model) to 65.0%. We interpret this as β-heterogeneity contributing +8.8% to the top 10% wealth share. Overall, discount factor heterogeneity does contribute positively to wealth inequality, but it is not the most important factor. Instead, differences in returns are crucial. Line 7 shuts down return differences across wealth levels ($r_t^X(\cdot) = 0$), line 8 return risk ($\sigma^X(\cdot) = 0$), while 6 combines the two modifications. Overall, differences in mean returns across wealth levels are far more important, though at the very top idiosyncratic return risk matters equally. Note that because model moments are highly non-linear as a function of parameters, individual modifications do not add up.

A striking feature of Table 2 is the fundamental importance of tax progressivity in keeping wealth inequality in check. Line 5 refers to a counterfactual that replaces the progressive income tax $\tau(\cdot)$ with a flat tax, such that aggregate tax revenues relative to output are unchanged. Wealth inequality is exploding. For example, the top 1% share increases from 27.4% to 89.2%. Why is tax progressivity so important? There are both partial- and general-equilibrium effects at work here. Starting with the latter, as we argued in Section 4.1, it is well known in the context of complete-markets models without heterogeneity in returns, discount factors, or earnings that progressivity in the tax rate on saving is a strong force toward long-run equality, whereas mere proportional taxes are consistent with any distribution of wealth as a steady-state equilibrium.[41] The mathematical intuition behind the force of progressivity

[41] Total wealth is of course pinned down so that the return to saving equal the discount rate, abstracting from consumption

is particularly clear in a simple case where the marginal tax rate is strictly increasing in wealth. Here, because all consumers face the same market rate under complete markets (and have the same discount rates and wage incomes), they also need to have the same net of tax return if their consumption levels are all constant (or growing at a common constant rate); hence they need to have the same wealth in the long run. This mechanism is still present in a more general model such as the present one, which has incomplete markets and differences in wages, returns, and discount rates, though with less long-run poignancy: a strictly increasing marginal tax rate is still consistent with long-run wealth inequality.

Turning to the partial-equilibrium analysis, note that the marginal saving propensity (out of initial-period assets) for a well-insured consumer with power utility is approximately $\beta(1 + r(1 - \tau'(y)))$ raised to a positive power, where $\tau'(y)$ is the consumer's current marginal tax rate.[42] This tax rate varies with the consumer's income, y, but it is persistent over time because income is persistent. Tax progressivity, therefore, generates persistent differences across consumers that act like persistent differences either in the consumers' after-tax rates-of-return, $r(1 - \tau'(y))$, or, equivalently, in consumers' discount factors. Consequently, *decreases* in progressivity have the same effect as increasing the dispersion in returns, a powerful force for generating higher wealth inequality.

Lines 2–4 in Table 2 document that earnings heterogeneity *reduces* wealth inequality in the benchmark model. Line 3 shuts down heterogeneity in the persistent component, line 4 likewise in the transitory component, while line 2 removes all earnings heterogeneity. Overall, both components reduce wealth inequality, though the strength of each of these channels depends somewhat on the particular wealth distribution statistic one looks at. To understand this finding, note first that without return or discount factor heterogeneity, earnings dispersion would contribute positively to wealth inequality. Then, why is the effect reversed in the benchmark model? As also noted in Section 4.1, heterogeneity in either discount factors (or returns) is a powerful force driving the wealth distribution apart: with permanently different discount rates and complete markets to insure against earnings risk, the most patient household would eventually hold all the economy's wealth. Earnings risk, then, is a friction, or glue, that keeps the distribution from flying apart altogether, as also in Becker (1980)'s work cited in Section 1. This risk operates especially strongly at the low end of the wealth distribution, where poorer consumers save to move away from borrowing constraints when earnings risk is larger.

In our model higher earnings risk also generates a thinner right tail in the wealth distribution because the resulting increase in aggregate precautionary savings drives down the equilibrium interest rate. This drop in the interest rate shifts the distribution of saving propensities to the left, particularly for the well-insured wealthy consumers for whom wage risk is largely immaterial and who therefore have essentially linear decision rules. As discussed in Section 5, the Pareto tail coefficient, ζ, is defined implicitly by the equation $\mathbb{E}[s^{\zeta}] = 1$, where s is the (asymptotic) marginal propensity to save out of wealth. As s falls for all discount-factor types, ζ must increase to compensate, i.e., the Pareto tail becomes thinner.[43]

growth.

[42] With $u(c) = \frac{c^{1-\sigma}-1}{1-\sigma}$, the power is $1/\sigma$.

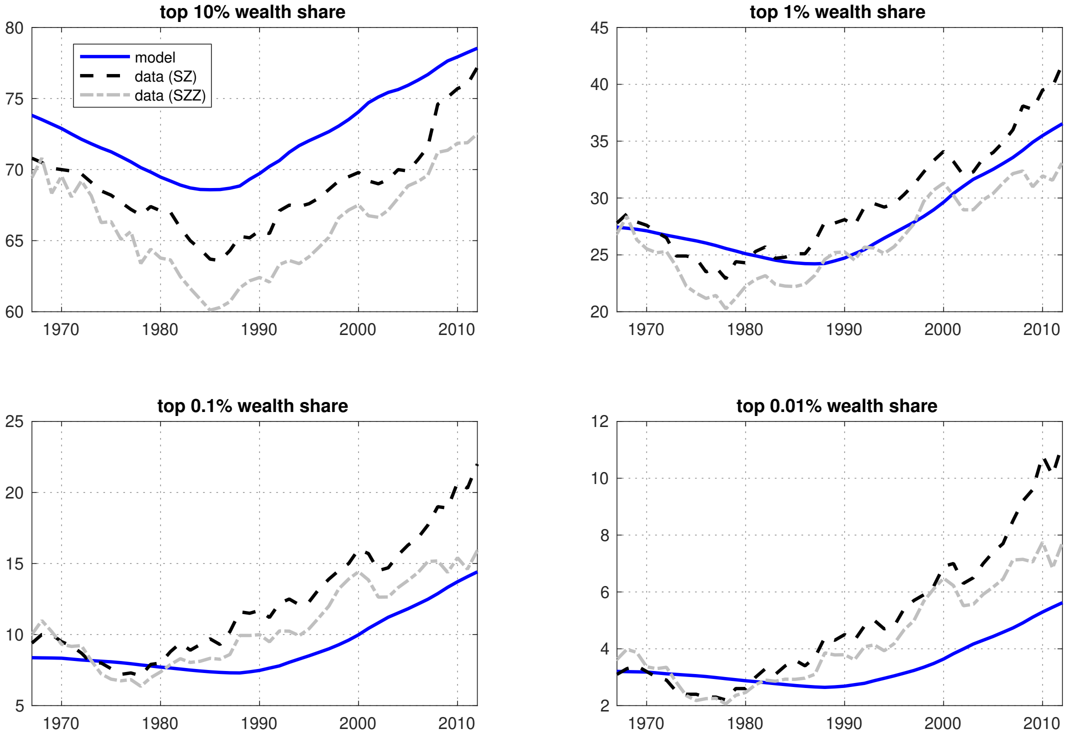
[43] Nirei & Aoki (2016) observe the same effect.

8 Results II: the evolution of the wealth distribution

In Section 7, we showed that our model framework, when properly calibrated, can replicate wealth heterogeneity, including the Pareto-shaped right tail, as well as other macroeconomic moments in the initial steady state. We proceed in this section to report on our second main result: the evolution of the wealth distribution in the benchmark model economy contrasted with the data. Subsequently, we employ counterfactual analysis in order to decompose those overall changes and identify the key drivers of movements in the wealth distribution.

8.1 Benchmark transition experiment

Figure 9: Top wealth shares in %, 1967–2012



Data sources: dashed black lines refer to Saez & Zucman (2016); dash-dotted gray lines refer to Smith et al. (2019).

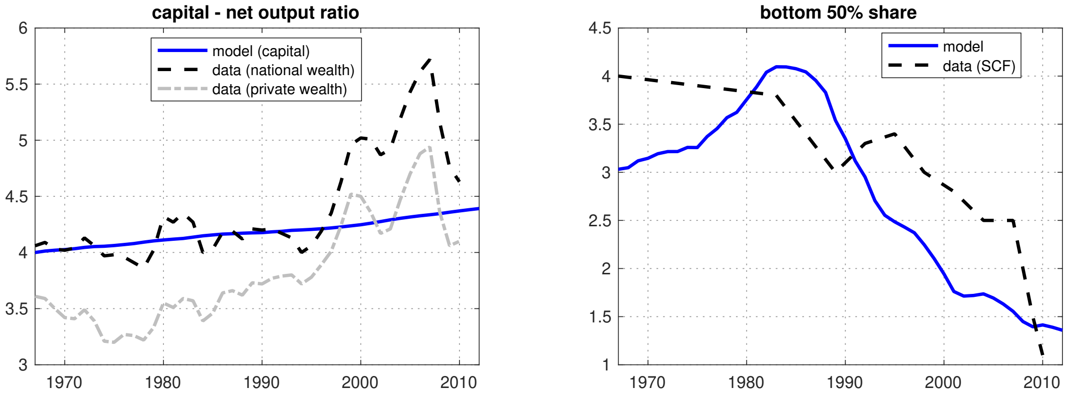
Figure 9 displays the evolution of top wealth shares in the model (solid blue line) compared to the data as measured by Saez & Zucman (2016) using the basic capitalization method (dashed black line; henceforth SZ). In addition, we augment the graphs with the estimates of Smith et al. (2019), who use a modified capitalization method approach that allows for return heterogeneity within asset classes (grey dash-dotted line; henceforth SZZ).[44] These shares display a pronounced U-shape, reaching the trough in the late 1970s to mid 1980s, followed by a sharp subsequent increase. The model economy matches both the initial decrease and the overall increase very well for the top 10% and the top 1%. Further in the

[44] As illustrated in Section 3, their findings align closely with those from the SCF.

tail, the model continues to capture the trend, though the increase is not quite as fast as estimated by SZ. In contrast, relative to SZZ, the model predicts a similar increase even for the very richest. As we discuss further in Section 9.3, the top wealth shares in the model economy continue to increase slowly over a long transition period before reaching the new steady state. This finding is consistent with Gabaix et al. (2016), who argue that the random growth mechanism that drives top wealth inequality tends to produce slow transitions (especially in the tails of the distribution).

Figure 10 displays the evolution of the capital-output ratio and of the bottom 50% wealth share. The model's implications for aggregate wealth are broadly in line with the data, thus showing a steady rise, ignoring shorter-run movements. The bottom 50% have lost a little over two thirds of their already small share of aggregate wealth; the model accounts for about two thirds of this decline in wealth.[45]

Figure 10: Capital-output ratio and bottom 50% share (in %), 1967–2012



Data sources: wealth-output ratios from Piketty & Zucman (2014); bottom 50% share from Kennickell (2011).

We also summarize the findings from our main experiment in Table 3. As illustrated above, both model and data show strong increases in wealth concentration at the top.

Looking at the top 10%, top 1%, and top 0.1% wealth share, the model explains between one half and three quarters of the cumulated increase in inequality as measured by SZ. On the other hand, relative to the more moderate estimates by SZZ, the model over-predicts the rise in concentration by about one third. In sum, while there is some ambiguity in the measurement of wealth inequality stemming from the lack of direct administrative data, the model is broadly in line with the data both qualitatively and quantitatively.

For the very richest, the model's performance is still qualitatively correct, but quantitatively it under-predicts the rise in wealth inequality. The model predicts an increase in the fraction of wealth held by the top 0.01% of about three quarters, whereas in the SZ data set the increase is even larger, by more than a factor of three. Relative to SZZ, the model explains about two thirds of the rise in wealth concentration.

Clearly, although—as suggested above—the basic capitalization method underlying the data may exaggerate the increases in wealth for the richest, this discrepancy is a major one unlikely to be solely due to mismeasurement and it does not appear that the present model is fully adequate for capturing the bulk

[45] The method of Saez & Zucman's unfortunately does not allow for a breakdown of the bottom 90% into subgroups.

Table 3: Change in top wealth shares

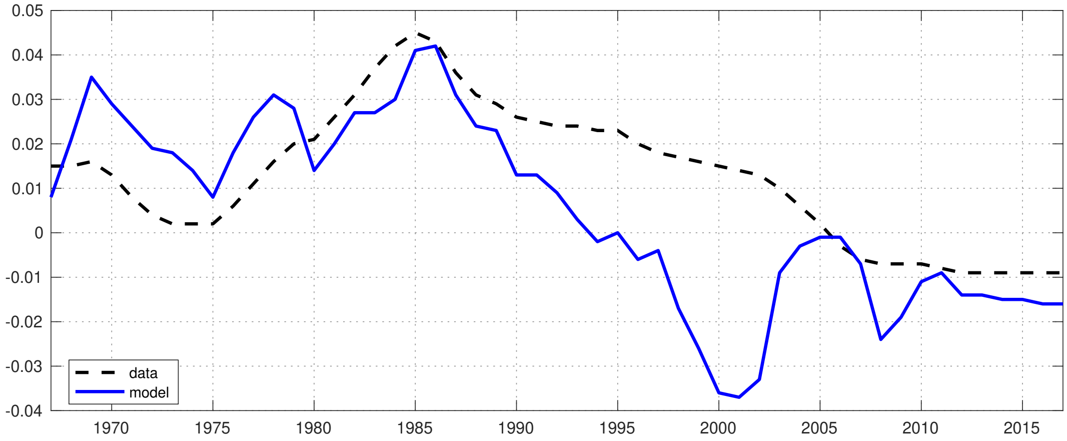
	Model			
	Top 10%	Top 1%	Top 0.1%	Top 0.01%
1967	73.8	27.4	8.4	3.2
2012	78.5	36.5	14.4	5.6
Change	4.7	9.1	6.0	2.4
Relative Change	6.4%	33.2%	72.2%	75.4%
	Data (Saez & Zucman)			
	Top 10%	Top 1%	Top 0.1%	Top 0.01%
1967	70.8	27.8	9.4	3.1
2012	77.2	41.8	22.0	11.2
Change	6.4	14.0	12.6	8.1
Relative Change	9.0%	50.4%	134.0%	261.3%
Fraction of Rel. Change Explained by Model	70.8%	65.9%	53.8%	28.9%
	Data (Smith, Zidar & Zwick)			
	Top 10%	Top 1%	Top 0.1%	Top 0.01%
1967	69.4	26.9	10.0	3.6
2012	72.5	33.2	15.9	7.7
Change	3.1	6.3	5.9	4.1
Relative Change	4.4%	23.5%	59.0%	113.3%
Fraction of Rel. Change Explained by Model	144.1%	141.4%	122.2%	66.6%

Data based on the capitalization method estimates of Saez & Zucman (2016), as well as Smith et al. (2019). Wealth shares are displayed in percentage points. For example, the top 1% controlled 27.8% of all wealth in 1967 according to SZ. By 2012, they controlled 41.8% of all wealth, an increase of 14 percentage points or 50.4% in relative terms. In the model, their share increased from 27.4% to 36.5%, an increase of 9.1 percentage points or 33.2%. Thus, the model explains a fraction $\frac{33.2}{50.4} = 65.9\%$ of the cumulative increase for this group.

of how much the richest have gained. There is an obvious remaining candidate explanation: idiosyncratic return volatility has gone up over time. The measures of idiosyncratic return volatilities (across wealth classes) cover only a short period of time, so we have no direct measure of movements in these volatilities over time. The increasing share of private equity in household portfolios—the flipside of which is a smaller and smaller share in publicly-traded stock—could be a source of increases in idiosyncratic return volatility. Moreover, Campbell, Lettau, Malkiel, & Xu (2001) provide evidence of increased volatility of individual stock returns (by about a factor of two, as measured by its standard deviation), so that if households held similarly undiversified portfolios throughout the period, their portfolio returns would indeed display increasing idiosyncratic volatility. We have not systematically examined this channel, as it involves much guesswork, but we have arbitrarily run an experiment where the volatility of private equity returns is doubled over the period and it indeed increases top wealth inequality by the end period quite significantly to levels comparable to those in the data. This is all suggestive but much more work is needed on this point.

Rates of return in model and data By construction, our model replicates the time paths of return premia on various asset classes perfectly. However, the overall level of returns is endogenous. Hence, the resulting equilibrium path for the risk-free rate is another model outcome that can be compared to the data to assess the performance of the model. Figure 11 shows that the model calibration, which targets the capital-output ratio, results in a real risk-free rate that is comparable to the data in the initial steady state.[46] Interestingly, the model reproduces almost the entire long-run decline. What drives the 2.4 percentage point decline in the model? The marginal product of capital falls by 0.9 percentage points in response to capital deepening, accounting for more than one third of the overall decline. The remainder is—in about equal parts—due to both rising return premia and rising wealth inequality: a higher fraction of aggregate wealth is held by rich agents, whose risky portfolio weights are higher; this depresses the risk-free rate in our setup.

Figure 11: The risk-free rate



The model is not doing as well in reproducing all medium-run swings. Around 2000, the model, unlike the data, generates a pronounced slump in the risk-free rate. Mechanically, this is because realized return premia on housing and equity were high. While in the data these high returns were largely driven by capital gains, in the model all capital income corresponds to the return on currently installed physical capital. Consequently, matching the return premia implies a counterfactually low risk-free rate.

8.2 Counterfactuals

Changes in four structural factors—earnings risk, top earnings inequality, tax progressivity, and excess returns—drive the transitional as well as long-run dynamics in the model economy. To assess which of these is the most important quantitatively, we conducted four experiments in which only one of the four structural factors is allowed to change, the other three being held constant instead at their values in 1967. Which of these changes is the main driver of increases in wealth inequality, particularly in the upper reaches of the distribution? As we shall see, the main driver of changes in the right tail of the wealth

[46] The data is heavily smoothed (using ten-year moving averages), and refers to the effective federal funds rate minus CPI inflation.

distribution is changes in taxes. Increases in earnings risk, on the other hand, *reduce* top wealth inequality, other things equal. Changes in return premia account in particular for the shorter-run dynamics.

Table 4 summarizes the results of the four experiments, quantifying how much each of the factors contributes to the changes in the wealth shares over the time period 1967–2012.[47]

Table 4: Fraction of change in wealth shares explained by model: decomposition by channel

	Bottom 50%	Top 10%	Top 1%	Top 0.1%	Top 0.01%
Taxes	0.49	1.57	1.15	0.72	0.36
Top Earnings Inequality	0.42	0.44	0.14	0.10	0.06
Earnings Risk	-0.04	-0.84	-0.21	-0.09	-0.05
Return Premia	-0.03	-0.58	-0.28	-0.13	-0.08
Combined	0.76	0.71	0.66	0.54	0.29

To understand the numbers in the table, focus on the share of total wealth held by the richest percentile. Saez & Zucman (2016) measure an increase in this share from 27.8% to 41.8% from 1967 to 2012. Over the same time period, allowing for changes only in earnings risk and keeping all other parameters fixed at their initial steady-state values, the model predicts a decrease from 27.4% to 24.5%. Changes in earnings risk therefore explain a fraction $\frac{24.5-27.4}{27.4}/\frac{41.8-27.8}{27.8} = -0.21$ of the actual change.[48] Again, the observed increases in earnings risk reduce inequality, moving it in the opposite direction from the observed changes! (Separate increases in either the persistent or transitory components of earnings risk also reduce inequality.) Instead, as can be seen for all the different distributional statistics, the main driver of the surge in wealth concentration is the changing U.S. tax system. The increase in top earnings inequality (parameterized by changes over time in the Pareto tail coefficient κ_t on labor income) has worked in the same direction, although the effect of this channel is much smaller. Changes in return premia have also dampened the increase in wealth concentration on net, in particular explaining the initial dip.

Why does an increase in earnings risk reduce wealth inequality? As explained in Sections 4.1 and 7.2, in the presence of return or discount factor heterogeneity, earnings risk can be viewed as the friction that prevents the wealth distribution from exploding. Facing higher earnings risk, consumers seek to increase precautionary savings, more so at the lower end of the wealth distribution. In addition, in general equilibrium the interest rate decreases, slowing down wealth accumulation at the top.

Why have changes in the tax system induced such large changes in wealth inequality? Note first that the average tax rate (i.e., aggregate tax revenues as a fraction of net output) in our model increases from 0.23 to 0.27 over the period 1967–2012. An increase in average taxes tends to reduce effective earnings risk (because the tax is multiplicative), increasing inequality for the same reason (but in the opposite direction) that the observed increases in (pre-tax) earnings risk reduce inequality. This effect, however, is a small one unless the average tax rate changes dramatically. Much more important quantitatively is the dramatic

[47] The dynamics are graphed in Figures 13 and 14 in the appendix.

[48] Note that the fractions generally do not add up to the fraction explained when feeding in all observed changes at the same time, as in our benchmark experiment. The remainder is due to interaction effects in general equilibrium.

decrease in tax progressivity, where even small changes have large effects on inequality, especially at the high end of the wealth distribution. As explained when discussing steady state inequality, tax progressivity effectively reduces dispersion in returns or discount factors, two powerful forces for driving the wealth distribution apart. Consequently, the observed decrease in progressivity triggered a large increase in wealth concentration. In Figure 15 in the appendix we decompose the effects of progressivity into a direct effect—the (mechanical) compression of after-tax resources induced by changes in progressivity, holding behavior fixed—and an indirect effect—the change in marginal saving propensities induced by changing progressivity, excluding its effects on the compression of after-tax resources—by showing the effects of the latter only and the effects of the former only, along with the full equilibrium response. Clearly, the direct effect is most important for the very richest and hence for changes in top wealth inequality.

Changes in return premia are key to explain the U-shape of wealth shares. In particular, the time series of private equity, primarily important for the rich, exhibits a pronounced U-shape (displayed earlier in Figure 8). Likewise, the average return on the U.S. stock market was quite low in the 1970s and 2000s. In contrast, house prices, particularly important for the middle class, have boomed until the Great Recession. Overall, changes in asset returns have reduced wealth inequality until about 1990, while contributing positively to increasing concentration subsequently.[49]

In sum, among the different drivers of wealth inequality considered in the benchmark experiment it is clear that decreasing tax progressivity is key: it spreads out the resources available to consume and invest and it increases the relative return of the rich on any given saving.

In a representative-agent model the increase in average taxes would lead to a decrease in the capital-to-output ratio in equilibrium, but it does not in our heterogeneous-agent model for three reasons. First, the (smallish) increase in average taxes does not offset the even larger increase in the riskiness of pre-tax earnings, leading to more precautionary savings in the aggregate. Second, decreasing tax progressivity increases the returns to saving, a particularly powerful force for the rich. Third, the increasingly "thick" right tail in earnings provides the rich (who tend to be those with high earnings) with additional resources for saving. These three forces combine to generate a fairly large increase in the ratio of capital to net output over the period 1967–2012.

If one looks at the wealth holdings of the bottom 50% of the population, the bulk of the decrease is again accounted for by the decrease in tax progressivity, as well as increases in top earnings inequality, while the movements in the aggregate capital-output ratio are mostly accounted for by the increase in earnings risk.[50] However, different measures can tell different stories. If one looks at the Gini coefficient for wealth *within* the bottom 50% or even within the bottom 90%, we find that the rise in earnings risk in our model does contribute positively to the increase in wealth inequality within this subgroup.

Let us now, finally, briefly compare our results to those in Kaymak & Poschke (2016). Their study emphasizes an increase in earnings inequality as a main driver of the increase in wealth inequality, but also finds the decline in tax progressivity to be important. As for their main finding, the effect of the increase

[49] Relatedly, Kuhn et al. (2019) assemble a long-run SCF data set for the post-war United States, and show in an accounting framework that differential exposure to asset price movements—due to differences in portfolio shares across wealth groups—accounts for a significant fraction of medium-run wealth inequality dynamics.

[50] Figure 14 in the appendix displays these results.

in earnings inequality in their model appears only after 1980, and after 1980 they consider only an increase in “top earnings” inequality—which is roughly similar to the top earnings inequality in our paper. Prior to 1980, they do not break down the effects on top wealth shares into a part that is due to changes in top earnings inequality and a part that is due to changes in earnings risk in the rest of the distribution. But, because wealth inequality is roughly constant before 1980 in their model, we conjecture that they go in opposite directions, just as they do in the present paper. Had we similarly considered only an increase in top earnings inequality after 1980, we would have obtained a positive effect of earnings inequality—as in their paper. Hence, overall, our models have similar predictions, with slightly different drivers, explaining the discrepancies in emphasis. As an important last remark, Kaymak & Poschke (2016) do not obtain the kind of U-shape in the evolution of inequality that we (and they) observe in the data; they do not consider the portfolio heterogeneity channel.

9 Extensions

We now look at a number of robustness exercises and extensions. First, we look at an aggregate production function with a non-unitary elasticity of substitution between capital and labor; our benchmark Cobb-Douglas (the unitary case) function does take a particular stand on the dynamics of the returns to capital. We find that this mechanisms does not appear very promising for understanding the data at hand.

We then weaken the consumers' ability to predict changes in their environment. In particular, in our benchmark experiment we assume that consumers in 1967 could predict the future paths of the tax schedule, the degree of idiosyncratic earnings risk, and even the return premia. These are of course strong assumptions, so it is interesting to compare this case to one with more limited abilities to predict. Here, our finding is that a model with entirely myopic expectations (the current policy/risk environment is expected to last forever) behaves almost like our benchmark environment.

Finally, we conclude the section with a cautious prediction for the long-run. Barring any future changes, the main message is that the adjustment process of the economy to the new steady state is far from over.

9.1 Robustness to the elasticity of substitution in production

The stability of the fraction of income accruing to labor, for a long time a central pillar of macroeconomic models, has recently been questioned. Karabarbounis & Neiman (2014b), among others, document a visible (though not large) decline in the labor share. Using a production function with a constant elasticity of substitution (CES), they estimate an elasticity of substitution between capital and labor of 1.25. To look into the possibility of a falling labor share, we use a standard CES production function,

$$F_{CES}(K_t, L) = A_{CES}\left(\alpha_{CES}K_t^{\frac{\sigma-1}{\sigma}} + (1-\alpha_{CES})L^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}, \tag{12}$$

where A_{CES} and α_{CES} are chosen such that the initial steady state is identical to the Cobb-Douglas benchmark. Over time, there is capital deepening, leading to a lower labor share because the elasticity of substitution is above one. We find, however, only very small differences as compared to the Cobb-Douglas benchmark (see Table 5).[51]

Table 5: Robustness to the input substitution elasticity and to myopia

		Top 10%	Top 1%	Top 0.1%	Top 0.01%	Bottom 50%	$\frac{K}{Y}$	r
1967		73.8	27.4	8.4	3.2	3.0	4.00	5.93
2013	Benchmark	78.9	37.1	14.8	5.8	1.3	4.40	5.11
	CES	78.6	36.7	14.6	5.7	1.3	4.45	5.20
	Myopia	76.9	34.9	14.1	5.7	1.4	4.42	5.07

Wealth shares and the interest rate r are reported in %. This table compares various statistics from the benchmark model transition to alternatives. In the benchmark transition experiment, the production technology is assumed to be Cobb-Douglas and agents have perfect foresight. The row labeled 'CES' reports results from a model with CES production technology. The row labeled 'Myopia' reports results from a transition experiment in which agents are completely myopic about the future, assuming present prices, returns, as well as the parameters of the earnings process and the tax schedule, will prevail.

Capital deepening leads to a smaller reaction of the interest rate, so the rise in the capital-output ratio is slightly larger in equilibrium and the Gini coefficient on gross income increases a small amount more (relative to the benchmark).[52] At the same time, we find that top wealth shares increase more slowly; unlike for the decline in tax progressivity, higher equilibrium interest rates induce more savings across the whole wealth distribution. In other words, at least over the time frame considered, the saving of the poor tends to be more elastic with respect to the interest rate than the saving of the rich. Overall, though, the message here is that the quantitative effects of considering a different elasticity of substitution are very small.

9.2 Robustness to agents' abilities to predict policy, risk, and returns

It is surely bold to assume that agents have perfect foresight on the entire path of the tax schedule, the parameters governing the earnings process, excess returns, and the resulting equilibrium prices. To gauge the sensitivity of our findings to this assumption, we computed the transitional dynamics under complete myopia, i.e., a polar opposite case in terms of agents' ability to predict. That is, in every period agents believe that the current environment will prevail forever and, accordingly, they are surprised to learn about their forecasting mistake in the subsequent period.[53] Table 5 shows the effects of myopia in the last row. Clearly, the differences are small. We conclude that the perfect-foresight assumption is not critically driving the results in the benchmark experiment. What is the reason for these results? One would perhaps particularly have guessed that being able to predict return movements would give rise to very different behaviors. Recall, however, that portfolio shares are hardwired and hence the household's ability to act on the foreseen changes in returns is limited. The same goes for the other factors: any

[51] Figure 16 in the appendix shows the time series.

[52] In addition, the gross labor share falls by about one percentage point over the period 1967–2012 in our model, though the net labor share actually rises a little. Karabarbounis & Neiman (2014a) report that since 1975 the gross labor share in the U.S. has fallen by about five percentage points and the net labor share by about two-and-a-half.

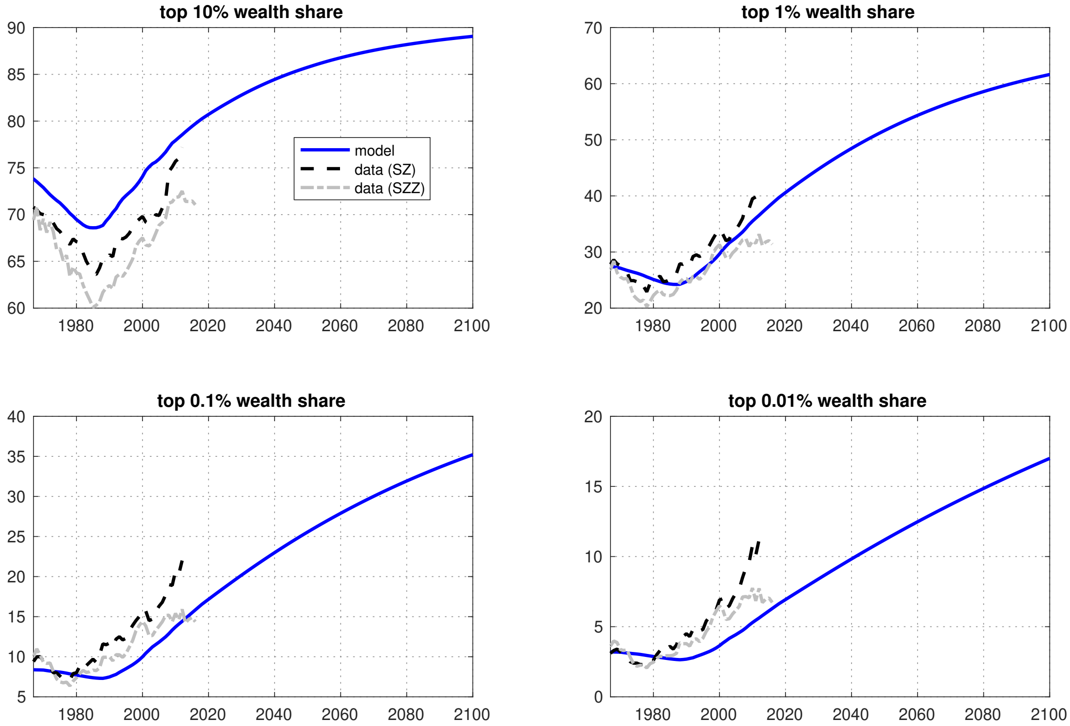
[53] See Section 4.3 for an exact description of how this experiment is conducted.

changes need to go through changes in saving rates, and these are rather robust.

9.3 The long run

We have focused so far on the transitional dynamics of the wealth distribution over the period 1967–2012, but what are the longer-run implications of the changes in earnings risk and, especially, tax progressivity that have occurred over this time period? In the calculations underlying these results, we have assumed no further changes in either earnings risk, taxes, or return premia after 2012.

Figure 12: Top wealth shares in %, long run



Data sources: dashed black lines refer to Saez & Zucman (2016); dash-dotted gray lines refer to Smith et al. (2019).

Figure 12 illustrates a striking prediction: the model suggests that the adjustment to the new fundamentals is far from completion and that wealth inequality is likely to rise even more. As pointed out before, the wealth distribution is a slow-moving object, especially in a setting with random growth in which the right tail of the wealth distribution is Pareto-shaped. Changes in fundamentals (such as the structure of taxes) that influence the consumption-savings decision differently for consumers with different wealth levels are bound, then, to have long-lasting effects. The contrast between the behavior of the wealth distribution over the transitional period and the eventual long-run steady-state wealth distribution (assuming an unchanged environment going forward) underscores the hazards of looking solely at steady states when attempting to quantify how fundamentals affect wealth inequality.

Of course, we urge caution in interpreting Figure 12 as a plain prediction for the future, because no doubt the economic environment will not remain unchanged going forward. Various exogenous impulses are possible (e.g., external forces affecting the U.S. interest rate, changes in demographics, and further change in earnings inequality). In addition, the model abstracts from plausible feedback mechanisms. For example, changes in wealth inequality could themselves, via the political process, lead to changes in the structure of taxes. In addition, accumulation of wealth in private equity by the richest could put downward pressure on its average return. Notwithstanding these points, the long-run analysis contained here does emphasize how powerfully tax progressivity can shape the wealth distribution, particularly in its right tail.

10 Concluding remarks

The determinants of wealth inequality, in particular its developments over the last half a century, are much-discussed recently and a number of new hypotheses have been put forth. This paper takes a "first-thing-first" perspective and asks what established quantitative theory predicts based on the behavior of a number of plausible, and observable, factors over the same period. We thus use a macroeconomic general-equilibrium model with heterogeneous agents—the Bewley-Huggett-Aiyagari setting—to examine more closely a set of candidate explanations for the increase in U.S. wealth inequality over the last forty or so years. The method we follow is thus to (i) independently measure changes in the environment, such as in the tax code, the earnings processes facing individuals, and their portfolio returns; (ii) feed these into the model assuming that the economy is in a steady state in 1967; (iii) examine the resulting wealth distribution path; and (iv) conduct counterfactuals. We find that the model generates a path for inequality that is quite close to that observed, the main exception being that the rise in inequality at the very top of the distribution is under-predicted if one takes the Saez-Zucman capitalization method as providing the right characterization of how the top wealth shares have evolved; compared to SCF data or the modified capitalization method of Smith et al. (2019), we instead over-predict the changes. The satisfactory performance of the model in predicting the overall path for wealth inequality notwithstanding, the first main contribution is the conclusion that the most important factor—by far—behind the long-run developments is the significant decline in tax progressivity that began in the late 1970s.

Declining tax progressivity, together with increasing earnings risk and higher earnings inequality amongst top earnings, can also account for the rise in the capital-to-net-output ratio and at least some of the decline in the (gross) labor share when the elasticity of substitution between capital and labor is larger than one as in Karabarbounis & Neiman (2014b). Our model thus provides an alternative to the central mechanism—declining growth rates—to which Piketty (2014) draws attention in attempting to connect these macroeconomic trends to rising inequality.

Our second major finding is that the key mechanism accounting for dynamics lies in heterogeneous portfolios across and within wealth groups, along with systematic return movements in the data.

Our third major finding, which is the one we discussed first in the paper, is the observation that in order to match wealth inequality in the beginning of the sample—which we do taking this year to represent a steady state—it is not necessary to add a "mop-up" explanation such as heterogeneous discount rates. Return heterogeneity is crucial here, giving a Pareto shape for the right wealth-distribution tail that significantly exceeds that for earnings.

Our findings merit several remarks. Although we find that tax progressivity has played a central role in increasing inequality, our model is designed primarily as a positive rather than a normative tool. To evaluate the pros and cons of, say, reversing the changes in tax progressivity, it is important to account for the distortions created by labor taxation; in the present setting, labor earnings are exogenous and taxation is levied jointly on all incomes. We do not think that the introduction of distortionary labor taxation would change the model's predictions for wealth inequality measurably, but it would be central for understanding the welfare consequences of tax changes. Further research contrasting the larger distortions of increased tax progressivity with the accompanying reductions in inequality seems very promising.

Our emphasis on differences in portfolios and portfolio returns between households is reminiscent of Piketty's stylized $r - g$ theory emphasizing the rate of return on assets, r, as an important determinant of the relative growth rates of wealth (including human wealth which grows at rate g) of the rich and the poor. The elaboration of this theory that we essentially propose is to attach less weight on g, to think of r as an after-tax return, and to recognize that r not only depends on household wealth (arising both from heterogeneous portfolio choices and progressive taxation) but also has a (stochastic) idiosyncratic component. This theory, moreover, mostly applies for the very richest; to understand the bulk and other side of the wealth distribution we side with Kaymak & Poschke (2016), who emphasize an increase in earnings inequality as a main driver of the increase in wealth inequality, but who also find the decline in tax progressivity to be important.

Regardless of one's normative views on wealth inequality, there are many reasons to care about its future course, as there are now many research contributions suggesting that the macroeconomy works quite differently when there is significant heterogeneity among consumers. This insight goes back at least as far as Krusell & Smith (1998) who showed that aggregate time series can depart significantly from permanent-income behavior in models in which wealth inequality matches the data. More recently, a growing body of research has demonstrated that both fiscal and monetary policy work differently too in models with proper microfoundations: for examples, see Heathcote (2005), McKay & Reis (2016), and Brinca, Holter, Krusell, & Malafry (2016) for fiscal policy, and Auclert (2019), McKay, Nakamura, & Steinsson (2016), and Kaplan, Moll, & Violante (2018) for monetary policy. The prediction from the present paper is that, barring reverses in the tax code, wealth inequality will go up even further, thus potentially strengthening the case for further research on the heterogeneous-agent approach to macroeconomics.

Finally, since so many of our findings rely on portfolio heterogeneity, we conclude by reiterating what we have stated repeatedly throughout the text: next, we need to understand households' portfolio choices better!

A Computational appendix

A.1 Dynamic programming problem

The consumers' dynamic programming problem is solved by value-function iteration using Carroll (2006)'s endogenous grid-point method (EGM) on a grid for cash-on-hand and the persistent idiosyncratic shocks (β, p).

Unlike in the plain Aiyagari (1994) model, the support of the ergodic wealth distribution is unbounded in this framework. We use a log-spaced grid with 100 points for cash-on-hand $(x_i)_{i=1}^{100}$ with a very large upper bound (one million times average wealth) to minimize the truncation error.[54] Cubic splines are used to interpolate the value function along the wealth dimension.

The grid for the persistent component of individual productivity $(p_j)_{j=1}^{17}$ is chosen to account for the long right tail in earnings. First, we chose the grid points as the 0.0001, 0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.925, 0.95, 0.975, 0.99, 0.999, ..., and 0.99999999 quantiles of the unconditional (i.e., cross-sectional) p-distribution (which is a normal). Second, we compute the corresponding grid in actual efficiency units of labor $(\psi(p_1), ..., \psi(p_{17}))$. Third, given that in the current period $p = p_j$ for $j = 1, ..., 17$, we use Gauss-Hermite quadrature to integrate over $p'|p$, the value of idiosyncratic productivity in the next period, when updating the value function. In doing so, we use linear interpolation in $\psi(p)$-space to evaluate the value function off the grid (the value function is much more non-linear in p-space than in $\psi(p)$-space).[55]

Regarding the discount factor, we choose the grid points $(\beta_m)_{m=1}^{15}$ as the Gauss-Hermite quadrature points of the unconditional (i.e., cross-sectional) β-distribution (this will turn out to be useful when integrating over the joint distribution to compute aggregate wealth). Again, when updating the value function, we integrate over $\beta'|\beta$ using Gauss-Hermite quadrature and linear interpolation in β-space.

In addition to the these three state variables, the setup requires numerical integration over the two idiosyncratic i.i.d. shocks to earnings ν' and capital returns η' (as they affect next period's cash-on-hand x'). As both shocks are normally distributed, we use Gauss-Hermite quadrature once again.

A.2 Computing the ergodic distribution

The focus on tiny population groups such as the top 0.01% of the wealth distribution implies that solving for the ergodic distribution directly is more efficient than simulating a large number of agents and applying the ergodic theorem. In doing so, simulation error is eliminated; instead one can directly control the numerical error by updating the distribution until convergence is reached.

Specifically, note that the EGM entails using a grid for assets $(a_i)_{i=1}^{100}$. Given p_j and β_m, saving a_i is

[54] Alternatively, given that the Pareto tail has stabilized at some $\bar{x}$, one could in principle also impute the distribution for $x > \bar{x}$. However, this did not turn out to be necessary as the log-spaced grid—which works well as the curvature of the value function is high only close to the borrowing constraint—allows for selecting a very large upper bound while keeping the number of grid points computationally feasible.

[55] Note that these linear interpolation coefficients can be pre-computed, resulting in a 17×17 - matrix w^p, where $w^p_{j,\cdot}$ are the integration weights for evaluating next period's value function on $(p_1, ..., p_{17})$ given that in the current period $p = p_j$.

optimal with cash-on-hand $x(a_i;p_j,\beta_m)$ that solves

$$\begin{aligned}\frac{\partial u(x(a_i;p_j,\beta_m)-a_i)}{\partial c} &= \beta_m \mathbb{E}\left[\left(1+\frac{\partial y'}{\partial a'}\left(1-\frac{\partial \tau(y')}{\partial y}\right)+\frac{\partial \tilde{y}'}{\partial a'}\left(1-\tilde{\tau}\right)\right)\frac{\partial V\left(x',p',\beta'\right)}{\partial x}|p_j,\beta_m\right],\\ &\text{where } x' = a_i + y' - \tau(y') + (1-\tilde{\tau})\tilde{y}' + T,\\ &\text{and } \frac{\partial y'}{\partial a'} = \left(\underline{r} + r^X(a_i) + \frac{\partial r^X(a_i)}{\partial a}a_i\right),\\ &\text{and } \frac{\partial \tilde{y}'}{\partial a'} = \left(\sigma^X(a_i)\eta' + \frac{\partial \sigma^X(a_i)}{\partial a}\eta' a_i\right).\end{aligned}$$

While the main advantage of the EGM is efficiency ($x(a_i;p_j,\beta_m)$ can be found without maximizing the right-hand side of the Bellman equation), it is also convenient that the savings function is already inverted. First, for all p_j,β_m,ν_q,η_h and for all a_i, $i=1,...,100$, there exists a unique level of asset holdings $a=s^{-1}(a_i;p_j,\beta_m,\nu_q,\eta_h)$ such that saving a_i is optimal.[56] Second, we define a finer grid for asset holdings $(k_i)_{i=1}^{1000}$ and interpolate (using a cubic spline) to find the inverse savings function $s^{-1}(k_i;p_j,\beta_m,\nu_q,\eta_h)$. Note that the borrowing constraint is binding for all $k \le s^{-1}(k_1;p_j,\beta_m,\nu_q,\eta_h)$. Finally, we can solve for the ergodic distribution $G(k_i;p_j,\beta_m) \equiv Prob(k \le k_i|p=p_j,\beta=\beta_m)$ at the grid points $(k_i)_{i=1}^{1000}$, $(p_j)_{j=1}^{17}$ and $(\beta_m)_{m=1}^{15}$. To simplify notation, we will denote by $G_{j,m}(k_i)$ this conditional cdf evaluated at grid points (p_j,β_m). This distribution has to satisfy

$$G_{j,m}(k_i) = \int_p\int_\beta\int_\nu\int_\eta G(s^{-1}(k_i;p,\beta,\nu,\eta);p,\beta)d\Gamma_\eta(\eta)d\Gamma_\nu(\nu)d\Gamma_\beta(\beta|\beta_m)d\Gamma_p(p|p_j). \tag{13}$$

Note that p_j and β_m are the realizations of the shock in period $t+1$ and the integration is over the shock values in period t. Nevertheless, e.g., $\Gamma_\beta(\beta|\beta_m)$ is the correct distribution as for any stationary Gaussian AR(1) process z_t the conditional random variables $z_t|z_{t+1}$ and $z_{t+1}|z_t$ have the same distribution.[57] Starting from some initial distribution $G^0_{j,m}(k_i)$ and using the short-hand notation $s^{-1}_{j,m,q,h}(k_i) = s^{-1}(k_i;p_j,\beta_m,\nu_q,\eta_h)$, we update until convergence according to

$$G^1_{j',m'}(k_i) = \sum_j w^p_{j',j}\sum_m w^\beta_{m',m}\sum_q w^\nu_q\sum_h w^\eta_h \hat{G}^0_{j,m}(s^{-1}_{j,m,q,h}(k_i)). \tag{14}$$

In (14), w^ν_q and w^η_h are the Gauss-Hermite quadrature weights for the transitory shocks ν and η (normalized to sum to one). The construction of the integration weights for the persistent shocks p and β is based on linear interpolation in $\psi(p)$- and β-space, respectively (see details below). $\hat{G}^0_{j,m}(\cdot)$ linearly interpolates $G^0_{j,m}(k_i)$ off the grid in the k-dimension.

Integration weights $w^p_{j,'j}$ and $w^\beta_{m',m}$. Consider the persistent earnings shock p. Conditional on its value in the next period being $p'=p_{j'}$ for some fixed $j' \in \{1,...,17\}$, the integration over the current

[56] $s^{-1}(a_i;p_j,\beta_m,\nu_q,\eta_h)$ is defined as the unique a that solves

$$x(a_i;p_j,\beta_m) = a + y - \tau(y) + (1-\tilde{\tau})\tilde{y} + T,$$

where $y = (\underline{r}+r^X(a))a + wl(p_j,\nu_q)$ and $\tilde{y} = \sigma^X(a)\eta_h a$.

[57] That is, the densities satisfy $f_{z_t|z_{t+1}}(x|y) = f_{z_{t+1}|z_t}(x|y)$.

period value p is with respect to the distribution of p, conditional on p', where $p|p' \sim N(\rho^P p' + (1 - \rho^P)\mu^P, \sigma^P)$. Gauss-Hermite quadrature, here with ten sample points, entails evaluating the function of interest $G(s^{-1}(k_i; p, \beta, \nu, \eta); p, \beta)$ at $(\tilde{p}_n)_{n=1}^{10}$, where $\tilde{p}_n = \rho^P p' + (1 - \rho^P)\mu^P + \sqrt{2}\sigma^P \tilde{x}_n$ and $(\tilde{x}_n)_{n=1}^{10}$ are the roots of the Hermite polynomial, and approximating the integral using the associated weights $(\tilde{w}_n)_{n=1}^{10}$ as

$$\approx \frac{1}{\sqrt{\pi}} \sum_{n=1}^{10} \tilde{w}_n G(s^{-1}(k_i; \tilde{p}_n, \beta, \nu, \eta); \tilde{p}_n, \beta).$$

Of course, $\tilde{p}_n$ will in general not lie on the p_j-grid, where the function value is known. Hence, we have to interpolate. Using linear interpolation, we can pre-compute the integration weights $(w^p_{j',j})_{j=1}^{17}$ we put on evaluating the function of interest at $(G(s^{-1}(k_i; p_j, \beta, \nu, \eta); p_j, \beta))_{j=1}^{17}$ in an efficient manner: for $n = 1, ..., 10$, locate $j(n)$ such that $p_{j(n)} \leq \tilde{p}_n \leq p_{j(n)+1}$ and compute the linear interpolation coefficient in $\psi(p)$-space λ_n as

$$\lambda_n = \frac{\psi(\tilde{p}_n) - \psi(p_{j(n)})}{\psi(p_{j(n)+1}) - \psi(p_{j(n)})}.$$

Then, looping over $n = 1, ..., 10$, add $(1 - \lambda_n)\frac{1}{\sqrt{\pi}}\tilde{w}_n$ to $w^p_{j',j(n)}$ and $\lambda_n \frac{1}{\sqrt{\pi}}\tilde{w}_n$ to $w^p_{j',j(n)+1}$. The construction of the integration weights for β is analogous, except that linear interpolation can be performed directly in β-space.

Computing moments of the distribution. For example, aggregate wealth is given by

$$K = \int_p \int_\beta \left(\int_k k dG(k|p, \beta) \right) f_p(p) f_\beta(\beta) dp d\beta,$$

where $f_p(\cdot)$ and $f_\beta(\cdot)$ are the unconditional (i.e., cross-sectional) normal densities of the persistent shocks p and β. We integrate numerically according to

$$\hat{K} = \sum_{j=1}^{17} \bar{w}^p_j \sum_{m=1}^{15} \bar{w}^\beta_m \left(k_1 G_{j,m}(k_1) + \sum_{i=2}^{1000} \frac{k_{i-1} + k_i}{2} \left(G_{j,m}(k_i) - G_{j,m}(k_{i-1}) \right) \right). \tag{15}$$

As the discount factor grid $(\beta_m)_{m=1}^{15}$ was chosen as the Gauss-Hermite sample points, we set $(\bar{w}^\beta_m)_{m=1}^{15}$ to be the associated Gauss-Hermite quadrature weights. Recall that the Pareto tail transformation of the persistent earnings component p prompted us to define a grid $(p_j)_{j=1}^{17}$ with a particular emphasis on the right tail. Hence, we (pre-)compute the integration weights $(\bar{w}^p_j)_{j=1}^{17}$ manually: (i) define a very fine equally spaced grid $(\hat{p}_n)_{n=1}^N$ (if, say, $N = 100,000$, this has to be carried out only once) that covers the coarser grid $(p_j)_{j=1}^{17}$; (ii) for all $n = 1, ..., N$, locate $j(n)$ and compute λ_n as above; (iii) looping over $n = 1, ..., N$, add $(1 - \lambda_n) f_p(\hat{p}_n)$ to $\bar{w}^p_{j(n)}$ and $\lambda_n f_p(\hat{p}_n)$ to $\bar{w}^p_{j(n)+1}$ ($f_p(\cdot)$ is the pdf of $p \sim N(\mu^P, \frac{\sigma^P)}{1-\rho^P})$; and (iv) finally, normalize such that $\sum_{j=1}^{17} \bar{w}^p_j = 1$.[58]

[58] Of course one could also use Gauss-Hermite quadrature here, as the corresponding weights and results coincide for all practical purposes.

A.3 Transition experiments

The perfect-foresight transition experiment is computationally straightforward. Given the calibrated initial steady state $(K^{\star}, \underline{r}^{\star}, T^{\star})$, the new steady state $(K^{\star\star}, \underline{r}^{\star\star}, T^{\star\star})$ is computed under the new exogenous environment. We then search for a fixed point in $(K_t, \underline{r}_t, T_t)_{t=t_0+1}^{t_1}$-space where $t_1 - t_0$ is chosen to be large enough that $(K_{t_1}, \underline{r}_{t_1}, T_{t_1}) \approx (K^{\star\star}, \underline{r}^{\star\star}, T^{\star\star})$. For each iteration, we first solve for the value functions and corresponding (inverse) savings decisions backwards and subsequently roll the distribution forward, as described in the previous sections for the steady state. Note that now the grids and integration weights for the earnings process components are time-varying.[59]

The myopic transition experiment is conceptually very different. Given a period t distribution $G_{j,m}^{t}(k_i)$ and savings decisions $s_{j,m,q,h}^{t}(k)$ (reflecting factor prices $\underline{r}_t$, w_t, transfers T_t and exogenous environment θ_t, all naively assumed to persist forever), $G_{j,m}^{t+1}(k_i)$ is obtained as in (14).[60] In turn, $G_{j,m}^{t+1}(k_i)$ and θ_{t+1} determine K_{t+1} (thus w_{t+1}), $\underline{r}_{t+1}$, and T_{t+1}. The surprised agents expect this new endogenous and exogenous environment to prevail forever and hence we solve the dynamic programming problem given this environment and accordingly obtain $s_{j,m,q,h}^{t+1}(k)$. Note that no fixed point problem has to be solved and the capital stock converges to the same new steady state as under perfect foresight. Theoretically, this strategy could give rise to oscillatory paths of capital. However, this turns out not to be the case in our application.

[59] In particular, as the variance of the innovation term of the persistent earnings component σ_t^P is time-varying, $p_{t|t+1}$ is no longer equal to $p_{t+1|t}$ in distribution (but still normal); hence the integration weights for the decision problem (forward-looking) and the cross-sectional distribution (backward-looking) differ.

[60] Again, the grids and integration weights for the earnings process components are time-varying.

B Additional figures

This section contains additional figures and results referred to in the main text.

Figure 13: Counterfactual top wealth shares in %, 1967–2012

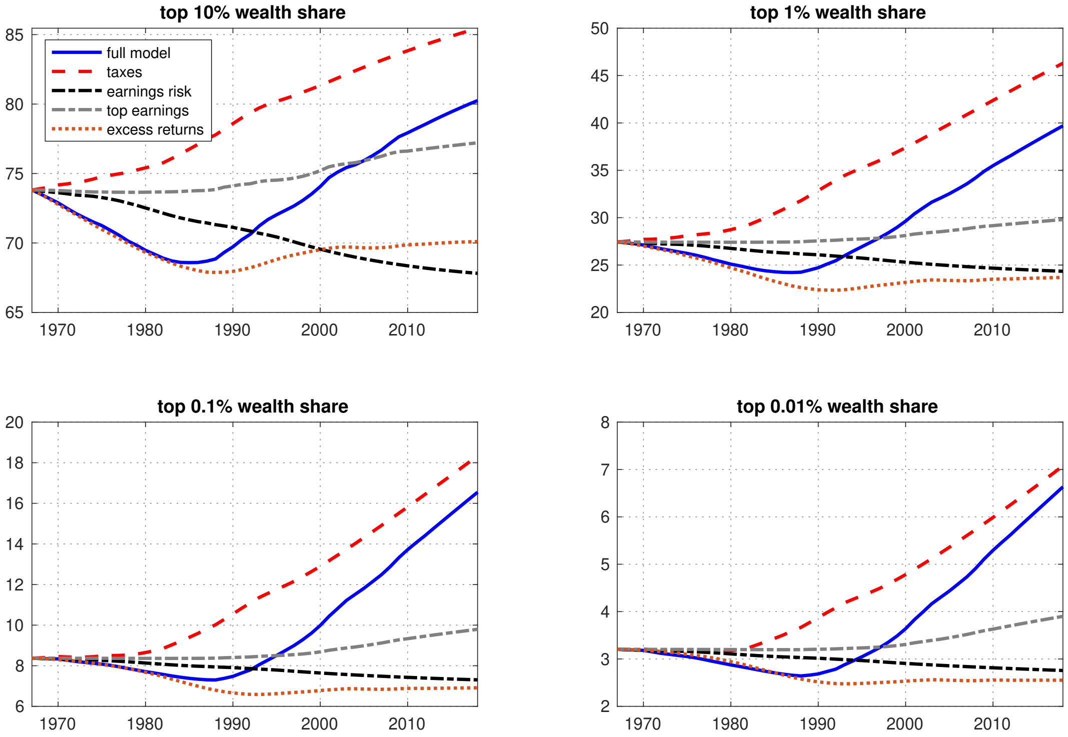


Figure 14: Counterfactual capital-output ratio and bottom 50% share (in %), 1967–2012

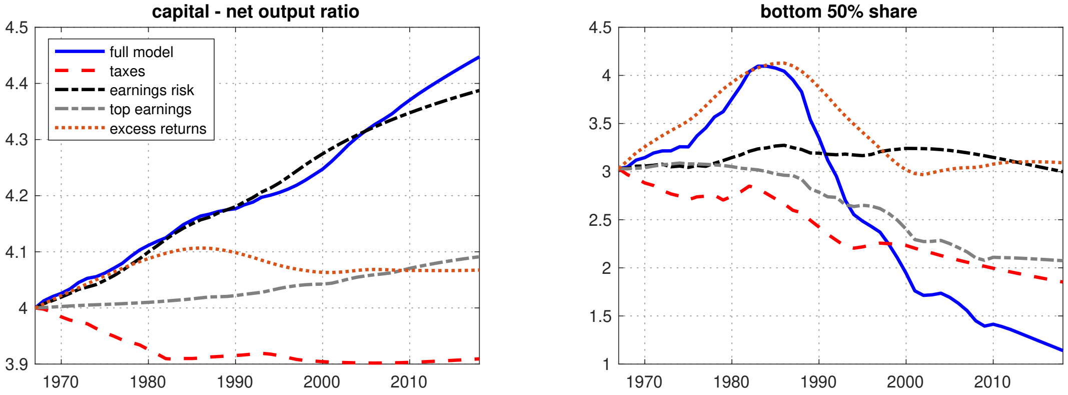


Figure 15: Tax-change decomposition: top wealth shares

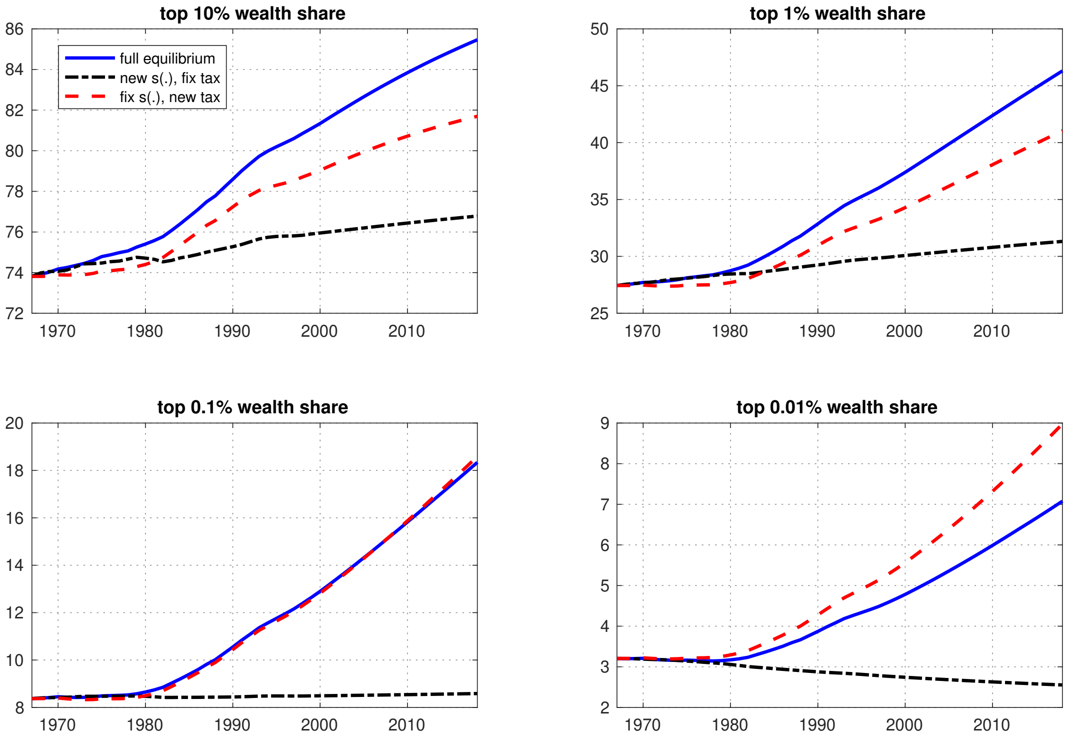
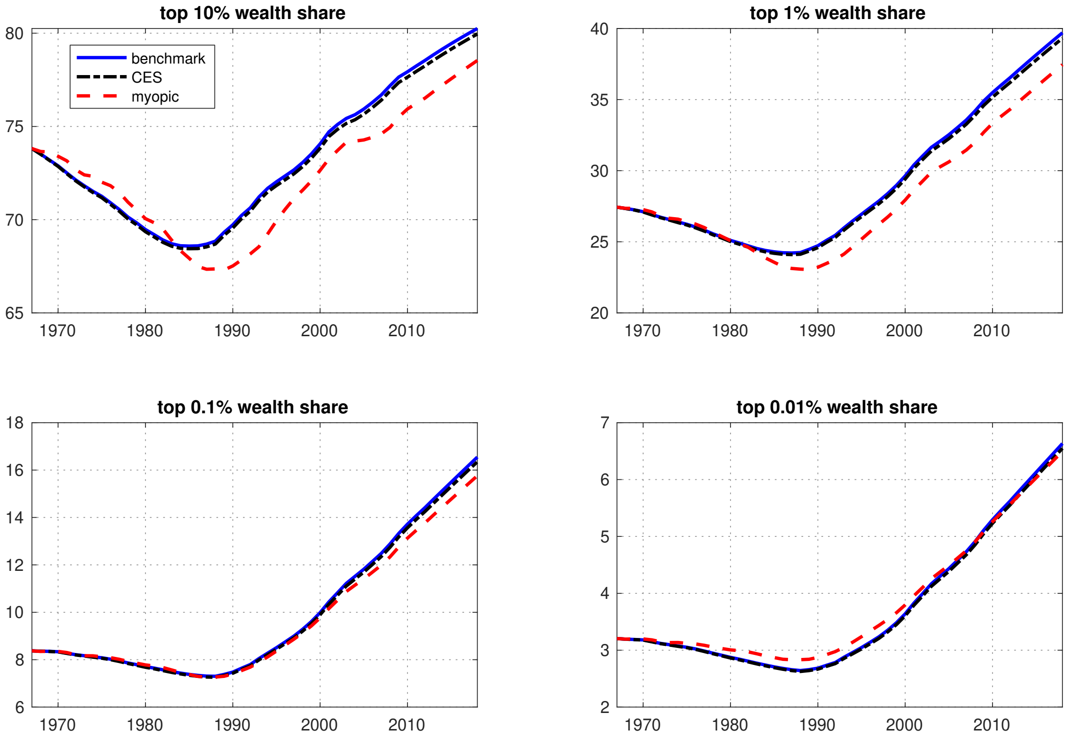


Figure 16: Robustness to myopia and CES production function with elasticity $\sigma = 1.25$



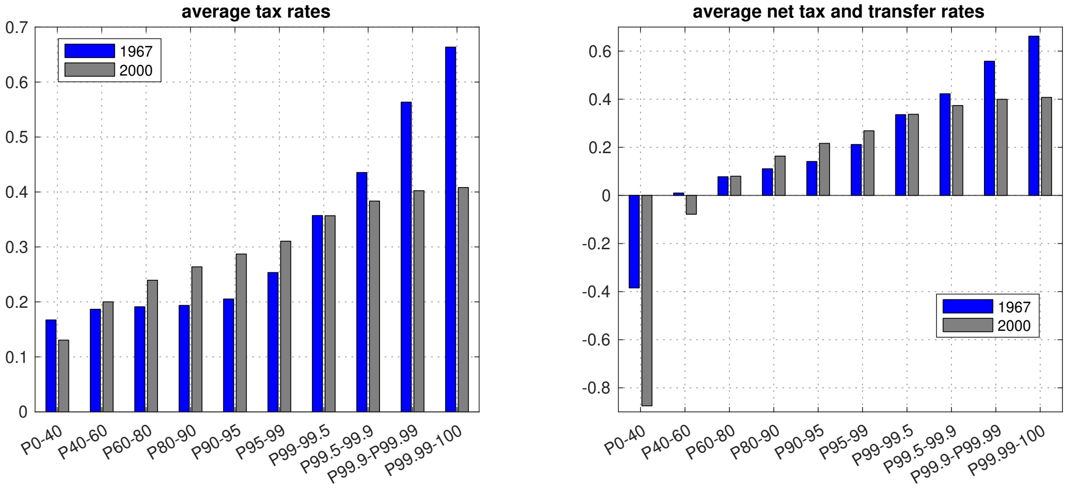
C Progressivity in the tax and transfer system

In the main text, we calibrate the tax function $\tau_t(\cdot)$, which is levied on gross income in the model, to data on gross tax rates. Since the model features a lump-sum transfer T_t, the implied net tax and transfer rate is more progressive. In particular, for a household j with income y_{jt} and gross tax rate $\tau_t(y_{jt})$, the net tax rate $n_t(y_{jt})$ is given by:

$$n_t(y_{jt}) = \tau_t(y_{jt}) - \frac{T_t}{y_{jt}}.$$

The left panel of Figure 17 shows gross tax rates in our model, which reflect the estimates of Piketty & Saez (2007). The right panel shows the combined net tax and transfer rate by income group in 1967 and 2000. Because (pre-tax) income inequality is substantial, the lump-sum nature of our transfer implies that the net tax and transfer system is much more progressive than the tax system exclusive of transfers. This property of U.S. taxes and transfers is highlighted by Auerbach, Kotlikoff, & Koehler (2019). Their estimates are not directly comparable, because we do not have a life-cycle in our model, but in agreement with their findings our model also produces quite substantial net subsidies at the lower end of the distribution. At the upper end, the transfer effectively vanishes in relative importance—gross and net rates coincide.

Figure 17: Progressivity in the tax and transfer system



D Excess return schedule details

Table 6: Details of excess return schedule

	P0-P40	P40-P50	P50-P60	P60-P70	P70-P80	P80-P90	P90-P95	P95-P97.5	P97.5-P99	P99-P99.5	P99.5-P99.9	P99.9-P99.99	Top 0.01%
fixed portfolio weights													
risk-free	0.722	0.412	0.248	0.182	0.156	0.134	0.115	0.102	0.090	0.079	0.071	0.051	0.029
housing	0.162	0.394	0.580	0.662	0.678	0.674	0.658	0.626	0.572	0.482	0.363	0.253	0.155
public equity	0.113	0.189	0.165	0.147	0.153	0.170	0.189	0.207	0.219	0.232	0.230	0.185	0.179
private equity	0.002	0.005	0.007	0.009	0.013	0.021	0.038	0.065	0.118	0.207	0.336	0.511	0.637
difference from aggregate return on asset class													
risk-free	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
housing	0.000	0.000	0.002	0.004	0.005	0.007	0.009	0.010	0.010	0.011	0.010	0.010	0.011
public equity	0.000	0.000	0.001	0.002	0.003	0.005	0.008	0.012	0.014	0.015	0.016	0.016	0.016
private equity	0.000	0.000	-0.019	-0.030	-0.054	-0.055	-0.049	-0.066	-0.064	-0.063	-0.063	-0.059	-0.060
standard deviation of return on asset class													
risk-free	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
housing	0.140	0.140	0.140	0.140	0.140	0.140	0.140	0.140	0.140	0.140	0.140	0.140	0.140
public equity	0.035	0.035	0.031	0.031	0.031	0.031	0.032	0.033	0.035	0.038	0.042	0.046	0.053
private equity	0.664	0.664	0.621	0.595	0.544	0.525	0.518	0.480	0.474	0.470	0.474	0.492	0.443
private equity (rescaled)	0.345	0.345	0.323	0.309	0.283	0.273	0.269	0.249	0.246	0.245	0.246	0.256	0.230
excess return schedule in 1967													
mean excess return	0.000	0.011	0.017	0.020	0.022	0.026	0.031	0.035	0.041	0.050	0.062	0.079	0.091
standard deviation	0.023	0.056	0.081	0.093	0.095	0.095	0.094	0.093	0.098	0.119	0.167	0.254	0.283
st. dev. (priv.equ. rescaled)	0.023	0.056	0.081	0.093	0.095	0.095	0.093	0.089	0.086	0.085	0.098	0.136	0.149

Portfolio weights by wealth group are adopted from Bach et al. (2019). The risk-free asset refers to bank account balances and money market funds. Housing refers to residential and commercial real estate. Public equity refers to risky financial assets in their data (financial assets minus risk-free assets). Private equity refers to shares of unlisted companies.

Differences from aggregate returns on a particular asset class by wealth group are also taken from Bach et al. (2019).

The standard deviation of returns by asset class and wealth group is also taken from Bach et al. (2019) (and corresponds to idiosyncratic risk only); except for housing, where we use an estimate for individual house price volatility in the U.S. from Piazzesi & Schneider (2016).

The resulting excess return schedule is then computed as described in the main text (using in addition aggregate excess returns by asset class from U.S. data).

E Comparison to U.S. portfolio and return data

Is the resulting excess return schedule, which we constructed using detailed information on Swedish households adopted from Bach et al. (2019), comparable to data on U.S. households? In an internet appendix, Bach et al. (2019) use U.S. Survey of Consumer Finance (SCF) data to show that overall return differentials by wealth group are remarkably similar in the U.S. and in Sweden. "For instance, the risk premium on gross wealth is 2.1% in the U.S. and 2.3% in Sweden on average for a household in the bottom decile; 3.7% in the U.S. and 3.8% in Sweden for the top 40%-50%; 5.3% in the U.S. and 5.6% in Sweden for the top 1%-0.5%. For the top 0.01%, however, a discrepancy arises: the risk premium on gross wealth is 6.6% in the U.S. and 7.5% in Sweden."[61] They argue that limited response rates and exclusion of the Forbes 400 are potentially an issue in SCF data (especially at the very top). Moreover, they document that when they impute excess returns in Swedish data, using only information that is comparably available in U.S. survey and aggregate data, return heterogeneity across wealth groups is substantially dampened. For this reason, and because it allows for within wealth group heterogeneity, we prefer calibrating our model using their high quality data.

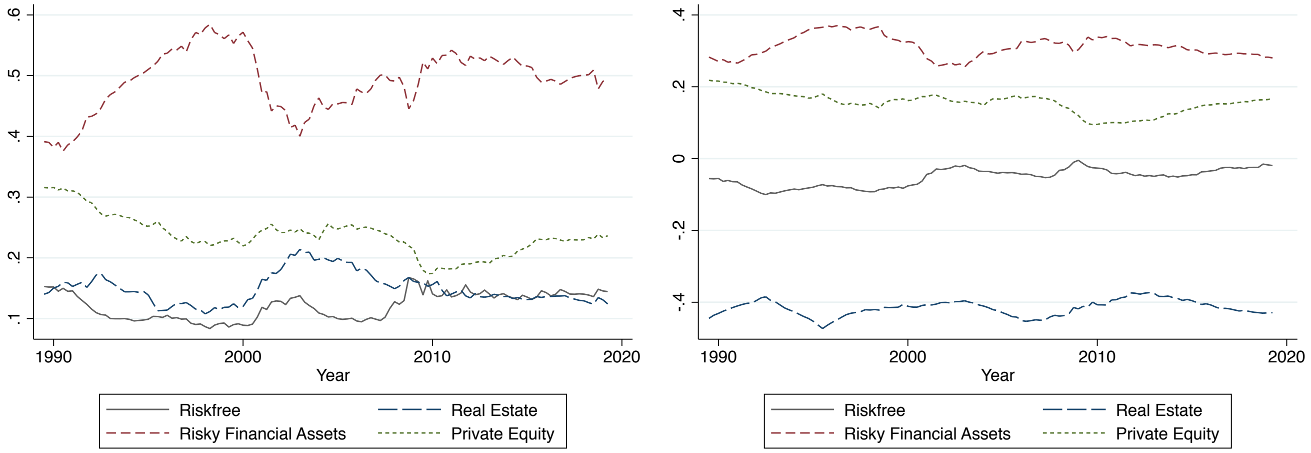
Relatedly, how restrictive is the assumption of constant portfolio weights over time? Figure 18 displays portfolio weights from the Distributional Financial Accounts (DFAs).[62] The left panel shows portfolio weights for the top 1% (no further disaggregation is provided within the top 1%). There is some time series variation. For example, the portfolio share of risky financial assets (including stocks) increases substantially in the 1990s. Presumably, households are not constantly re-balancing their portfolios (ceteris paribus, a boom in stock prices must necessarily increase the aggregate portfolio weight of public equity, and by extension the portfolio weight of most households in a value-weighted sense). However, what matters in our model are primarily differences in portfolio shares between wealth groups.[63] The right panel of Figure 18 plots differences between the portfolio shares of the top 1% and the middle class (P50-90). These differential portfolio shares are relatively constant over time. It is in this sense that we are confident that our assumption of constant portfolio weights does not substantially affect our quantitative findings.

[61] See `https://sites.google.com/site/laurentbach/AppendixRichPickings.pdf`, p.52–54.

[62] The DFAs are a new data product harmonizing SCF household data with aggregate data from the Financial Accounts. See Batty, Briggs, Pence, Smith, & Volz (2019). We aggregate assets in asset classes that are comparable to the classification in Bach et al. (2019), which we use to calibrate our model. The risk-free class comprises of checkable deposits and currency, time deposits and short-term investments, and money market funds. The risky financial asset class consists of corporate equities and mutual fund shares (69.2%), debt securities (20.1%), other loans and life insurance reserves (10.7%). Private equity refers to equity in non-corporate business. We exclude consumer durables and pension entitlements.

[63] To see this, consider a uniform increase in the risky asset share (and a corresponding decrease in the risk-free asset share). Provided the risk premium $\bar{r}_{c,t}$ is positive, the mean excess return $r_t^X(a)$ increases for all wealth levels. In turn, market clearing in our setup requires a corresponding decrease in the common component $\underline{r}_t$, such that the expected return $\underline{r}_t + r_t^X(a)$ is unchanged. Since our setup also features heterogeneity within asset classes across wealth groups (and within wealth groups), such a uniform shift in portfolio weights would nevertheless have a small effect on returns.

Figure 18: U.S. household portfolio shares 1989–2019



(a) Top 1% portfolio shares

(b) Top 1% − middle class portfolio shares

Source: U.S. Distributional Financial Accounts.

References

Acemoglu, D. (2002). Technical Change, Inequality, and the Labor Market. *Journal of Economic Literature*, *40*(1), 7–72.

Aiyagari, S. R. (1994). Uninsured Idiosyncratic Risk and Aggregate Saving. *The Quarterly Journal of Economics*, *109*(3), pp. 659–684.

Aoki, S., & Nirei, M. (2017). Zipf's law, pareto's law, and the evolution of top incomes in the united states. *American Economic Journal: Macroeconomics*, *9*(3), 36–71.
URL http://www.aeaweb.org/articles?id=10.1257/mac.20150051

Auclert, A. (2019). Monetary policy and the redistribution channel. *American Economic Review*, *109*(6), 2333–67.
URL http://www.aeaweb.org/articles?id=10.1257/aer.20160137

Auerbach, A. J., Kotlikoff, L. J., & Koehler, D. (2019). U.S. Inequality and Fiscal Progressivity—An Intragenerational Accounting. Working paper.

Bach, L., Calvet, L. E., & Sodini, P. (2019). Rich Pickings? Risk, Return, and Skill in the Portfolios of the Wealthy. Working paper.

Batty, M., Briggs, J., Pence, K., Smith, P., & Volz, A. (2019). The Distributional Financial Accounts. Feds notes, Board of Governors of the Federal Reserve System.
URL https://doi.org/10.17016/2380-7172.2436

Becker, R. A. (1980). On the Long-Run Steady State in a Simple Dynamic Model of Equilibrium with Heterogeneous Households. *The Quarterly Journal of Economics*, *95*(2), 375–382.

Benhabib, J., Bisin, A., & Luo, M. (2017). Earnings inequality and other determinants of wealth inequality. *American Economic Review*, *107*(5), 593–97.
URL http://www.aeaweb.org/articles?id=10.1257/aer.p20171005

Benhabib, J., Bisin, A., & Luo, M. (2019). Wealth distribution and social mobility in the us: A quantitative approach. *American Economic Review*, *109*(5), 1623–47.
URL http://www.aeaweb.org/articles?id=10.1257/aer.20151684

Benhabib, J., Bisin, A., & Zhu, S. (2011). The Distribution of Wealth and Fiscal Policy in Economies With Finitely Lived Agents. *Econometrica*, *79*(1), 123–157.

Benhabib, J., Bisin, A., & Zhu, S. (2015). The Wealth Distribution in Bewley Economies with Capital Income Risk. *Journal of Economic Theory*, *159, Part A*, 489 – 515.
URL http://www.sciencedirect.com/science/article/pii/S0022053115001362

Bewley, T. (undated). Interest Bearing Money and the Equilibrium Stock of Capital. Manuscript.

Bricker, J., Henriques, A., Krimmel, J., & Sabelhaus, J. (2016). Measuring Income and Wealth at the Top Using Administrative and Survey Data. *Brookings Papers on Economic Activity*, *Spring 2016*, 261–331.

Brinca, P., Holter, H., Krusell, P., & Malafry, L. (2016). Fiscal Multipliers in the 21st Century. *Journal of Monetary Economics*, *77*, 53–69.

Cagetti, M., & De Nardi, M. (2006). Entrepreneurship, Frictions, and Wealth. *Journal of Political Economy*, *114*(5), 835–870.

Cagetti, M., & De Nardi, M. (2009). Estate Taxation, Entrepreneurship, and Wealth. *American Economic Review*, *99*(1), 85–111.

Campanale, C. (2007). Increasing Returns to Savings and Wealth Inequality. *Review of Economic Dynamics*, *10*(4), 646–675.
URL `https://ideas.repec.org/a/red/issued/04-102.html`

Campbell, J. Y., Lettau, M., Malkiel, B. G., & Xu, Y. (2001). Have individual stocks become more volatile? an empirical exploration of idiosyncratic risk. *The Journal of Finance*, *56*(1), 1–43.
URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/0022-1082.00318`

Carroll, C. D. (2006). The Method of Endogenous Gridpoints for Solving Dynamic Stochastic Optimization Problems. *Economics Letters*, *91*(3), 312–320.

Carroll, C. D. (2012). Theoretical Foundations of Buffer Stock Saving. Working paper.

Carroll, C. D., & Kimball, M. S. (1996). On the Concavity of the Consumption Function. *Econometrica*, *64*(4), 981–92.

Castañeda, A., Días-Giménez, J., & Ríos-Rull, J.-V. (2003). Accounting for the U.S. Earnings and Wealth Inequality. *Journal of Political Economy*, *111*(4), 818–857.

CBO (2015). The Budget and Economic Outlook: 2015 to 2025. Tech. rep., Congressional Budget Office.

Chatterjee, S. (1994). Transitional dynamics and the distribution of wealth in a neoclassical growth model. *Journal of Public Economics*, *54*(1), 97–119.
URL `https://ideas.repec.org/a/eee/pubeco/v54y1994i1p97-119.html`

Cronqvist, H., & Siegel, S. (2015). The origins of savings behavior. *Journal of Political Economy*, *123*(1), 123–169.
URL `https://doi.org/10.1086/679284`

Fagereng, A., Guiso, L., Malacrino, D., & Pistaferri, L. (2020). Heterogeneity and Persistence in Returns to Wealth. *Econometrica*, *88*(1), 115–170.
URL `https://ideas.repec.org/a/wly/emetrp/v88y2020i1p115-170.html`

Gabaix, X. (2009). Power Laws in Economics and Finance. *Annual Review of Economics*, *1*(1), 255–294.

Gabaix, X., Lasry, J.-M., Lions, P.-L., & Moll, B. (2016). The Dynamics of Inequality. *Econometrica*, *84*(6), 2071–2111.
URL `http://dx.doi.org/10.3982/ECTA13569`

Guerrieri, V., & Lorenzoni, G. (2017). Credit Crises, Precautionary Savings, and the Liquidity Trap. *The Quarterly Journal of Economics*, *132*(3), 1427–1467.
URL `https://ideas.repec.org/a/oup/qjecon/v132y2017i3p1427-1467..html`

Heathcote, J. (2005). Fiscal policy with heterogeneous agents and incomplete markets. *Review of Economic Studies*, *72*(1), 161–188.

Heathcote, J., Storesletten, K., & Violante, G. L. (2010). The Macroeconomic Implications of Rising Wage Inequality in the United States. *Journal of Political Economy*, *118*(4), 681–722.

Hornstein, A., Krusell, P., & Violante, G. (2005). The Effects of Technical Change on Labor Market Inequalities . In P. Aghion, & S. Durlauf (Eds.) *Handbook of Economic Growth*, vol. 1 of *Handbook of Economic Growth*, (pp. 1275 – 1370). Elsevier.

Huggett, M. (1993). The Risk-Free Rate in Heterogeneous-Agent Incomplete-Insurance Economies. *Journal of Economic Dynamics and Control*, *17*(5-6), 953–969.

Jordà, Ò., Knoll, K., Kuvshinov, D., Schularick, M., & Taylor, A. M. (2019). The Rate of Return on Everything, 1870–2015. *The Quarterly Journal of Economics*, *134*(3), 1225–1298.
URL `https://ideas.repec.org/a/oup/qjecon/v134y2019i3p1225-1298..html`

Kaplan, G., Moll, B., & Violante, G. L. (2018). Monetary policy according to hank. *American Economic Review*, *108*(3), 697–743.
URL `http://www.aeaweb.org/articles?id=10.1257/aer.20160042`

Karabarbounis, L., & Neiman, B. (2014a). Capital Depreciation and Labor Shares Around the World: Measurement and Implications. Working paper.

Karabarbounis, L., & Neiman, B. (2014b). The Global Decline of the Labor Share. *The Quarterly Journal of Economics*, *129*(1), 61–103.

Kartashova, K. (2014). Private equity premium puzzle revisited. *American Economic Review*, *104*(10), 3297–3334.
URL `http://www.aeaweb.org/articles?id=10.1257/aer.104.10.3297`

Katz, L. F., & Murphy, K. M. (1992). Changes in Relative Wages, 1963-1987: Supply and Demand Factors. *The Quarterly Journal of Economics*, *107*(1), 35–78.

Kaymak, B., & Poschke, M. (2016). The Evolution of Wealth Inequality over Half a Century: The Role of Taxes, Transfers and Technology. *Journal of Monetary Economics*, *77*(C), 1–25.

Kennickell, A. B. (2011). Tossed and Turned: Wealth Dynamics of U.S. Households 2007-2009. Finance and Economics Discussion Series 2011-51, Board of Governors of the Federal Reserve System.

Kesten, H. (1973). Random Difference Equations and Renewal Theory for Products of Random Matrices. *Acta Mathematica*, *131*(1), 207–248.

Kopczuk, W. (2015). What Do We Know about the Evolution of Top Wealth Shares in the United States? *Journal of Economic Perspectives*, *29*(1), 47–66.

Kopczuk, W., & Saez, E. (2004). Top Wealth Shares in the United States, 1916-2000: Evidence from Estate Tax Returns. *National Tax Journal*, *2, Part 2*, 445–487.

Krusell, P., Mukoyama, T., Şahin, A., & Smith, A. A., Jr. (2009). Revisiting the Welfare Effects of Eliminating Business Cycles. *Review of Economic Dynamics*, *12*, 393–404.

Krusell, P., & Smith, A. A., Jr. (1998). Income and Wealth Heterogeneity in the Macroeconomy. *Journal of Political Economy*, *106*(5), 867–896.

Krusell, P., & Smith, A. A., Jr. (2006). Quantitative Macroeconomic Models with Heterogeneous Agents. In R. Blundell, W. Newey, & T. Persson (Eds.) *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, Econometric Society Monographs, 41*, (pp. 298–340). Cambridge University Press.

Krusell, P., & Smith, A. A., Jr. (2015). Is Piketty's "Second Law of Capitalism" Fundamental? *Journal of Political Economy*, *123*(4), 725–748.

Kuhn, M., Schularick, M., & Steins, U. I. (2019). Income and Wealth Inequality in America, 1949-2016. *Journal of Political Economy*, *forthcoming*.

McKay, A., Nakamura, E., & Steinsson, J. (2016). The Power of Forward Guidance Revisited. *American Economic Review*, *106*(10), 3133–3158.

McKay, A., & Reis, R. (2016). The Role of Automatic Stabilizers in the U.S. Business Cycle. *Econometrica*, *84*(1), 141–194.

Mehra, R., & Prescott, E. C. (1985). The equity premium: A puzzle. *Journal of Monetary Economics*, *15*(2), 145 – 161.
URL http://www.sciencedirect.com/science/article/pii/0304393285900613

Nirei, M., & Aoki, S. (2016). Pareto Distribution of Income in Neoclassical Growth Models. *Review of Economic Dynamics*, *20*(1), 25–42.

Piazzesi, M., & Schneider, M. (2016). *Housing and Macroeconomics*, vol. 2 of *Handbook of Macroeconomics*, chap. 0, (pp. 1547–1640). Elsevier.
URL https://ideas.repec.org/h/eee/macchp/v2-1547.html

Piketty, T. (1995). Social Mobility and Redistributive Politics. *The Quarterly Journal of Economics*, *110*(3), 551–84.

Piketty, T. (1997). The Dynamics of the Wealth Distribution and the Interest Rate with Credit Rationing. *Review of Economic Studies*, *64*, 173–189.

Piketty, T. (2003). Income inequality in france, 1901–1998. *Journal of political economy*, *111*(5), 1004–1042.

Piketty, T. (2014). *Capital in the Twenty-First Century*. Translated by Arthur Goldhammer. Cambridge, MA: Belknap.

Piketty, T., & Saez, E. (2003). Income Inequality in the United States, 1913-1998. *The Quarterly Journal of Economics*, *118*(1), 1–41.

Piketty, T., & Saez, E. (2007). How Progressive is the U.S. Federal Tax System? A Historical and International Perspective. *Journal of Economic Perspectives*, *21*(1), 3–24.

Piketty, T., Saez, E., & Zucman, G. (2018). Distributional National Accounts: Methods and Estimates for the United States. *The Quarterly Journal of Economics*, *133*(2), 553–609.
URL `https://ideas.repec.org/a/oup/qjecon/v133y2018i2p553-609..html`

Piketty, T., & Zucman, G. (2014). Capital is Back: Wealth-Income Ratios in Rich Countries 1700-2010. *The Quarterly Journal of Economics*, *129*(3), 1255–1310.

Piketty, T., & Zucman, G. (2015). Wealth and inheritance in the long run (chapter 15). In A. B. Atkinson, & F. Bourguignon (Eds.) *Handbook of Income Distribution*, vol. 2 of *Handbook of Income Distribution*, (pp. 1303 – 1368). Elsevier.

Quadrini, V. (2000). Entrepreneurship, Saving, and Social Mobility. *Review of Economic Dynamics*, *3*(1), 1–40.

Quadrini, V., & Rios-Rull, J.-V. (2015). Inequality in macroeconomics (chapter 14). In A. B. Atkinson, & F. Bourguignon (Eds.) *Handbook of Income Distribution*, vol. 2 of *Handbook of Income Distribution*, (pp. 1229 – 1302). Elsevier.

Saez, E., & Zucman, G. (2016). Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data. *Quarterly Journal of Economics*, *2*, 519–578.

Saez, E., & Zucman, G. (2019). Progressive Wealth Taxation. *Brookings Paper on Economic Activity*.
URL `https://www.brookings.edu/bpea-articles/progressive-wealth-taxation/`

Smith, M., Zidar, O., & Zwick, E. (2019). Top Wealth in the United States: New Estimates and Implications for Taxing the Rich. Working paper.
URL `http://ericzwick.com/wealth/wealth.pdf`

Stachurski, J., & Toda, A. A. (2019). An impossibility theorem for wealth in heterogeneous-agent models with limited heterogeneity. *Journal of Economic Theory*, *182*(C), 1–24.
URL `https://ideas.repec.org/a/eee/jetheo/v182y2019icp1-24.html`

Stiglitz, J. E. (1969). Distribution of Income and Wealth Among Individuals. *Econometrica*, *37*(3), 382–397.

Toda, A. A. (2014). Incomplete market dynamics and cross-sectional distributions. *Journal of Economic Theory*, *154*(C), 310–348.
URL `https://ideas.repec.org/a/eee/jetheo/v154y2014icp310-348.html`

Toda, A. A. (2018). Wealth distribution with random discount factors. *Journal of Monetary Economics*.
URL `http://www.sciencedirect.com/science/article/pii/S0304393218305592`

U.S. Department of the Treasury (2016). Office of Tax Analysis: Taxes Paid on Capital Gains for Returns with Positive Net Capital Gains, 1954-2014. Tech. rep.