Essays on Structural Change and Macroeconomic Outcomes

Chenchuan Shi

Nuffield College

University of Oxford

Submitted in fulfillment of the requirements
for the degree of

the Doctor of Philosophy in Economics

April 2026

Abstract

This thesis, Essays on Structural Change and Macroeconomic Outcomes, studies how demographic, organizational, and technological change shape macroeconomic outcomes. It combines structural models with micro data to uncover the mechanisms linking individual behavior to aggregate dynamics.

The first chapter analyzes how population aging interacts with ambiguity aversion and impacts on portfolio choice. With rising life expectancy, wealthier and older individuals become endogenously less pessimistic about risky assets, while younger households become more pessimistic, leading to increasing intergenerational inequality. Survey data support this mechanism, and model projections imply a 26% increase in the age gradient of risky asset holdings by 2100.

The second chapter studies the macroeconomic impact of remote work. We build a model where firms trade off productivity losses against cost savings associated with remote work. We show that the expansion of remote work raises profitability and encourages firm entry, but shifts the distribution toward smaller firms. Entry barriers are therefore key in determining aggregate productivity and overall welfare effects.

The third chapter studies how the generality of knowledge, its applicability across technologies, shapes firms' innovation strategies, market structure, and growth. I build an endogenous growth model and show that (1) market leaders prefer firm-specific R&D, and followers rely on general R&D to catch up; (2) the gap in innovation generality between them follows a U-shaped pattern with market concentration. Evidence from variation in state-level non-compete enforceability supports the model's spillover mechanism. The findings highlight a novel growth policy: promoting general R&D, especially among leaders, is growth-enhancing.

Acknowledgment

I express my deep gratitude to my supervisors, Martin Ellison and Petr Sedláček for their guidance and support throughout the journey of my PhD study.

I gratefully acknowledge the Oxford-Adrian Swire Graduate Scholarship, Department of Economics Doctoral Bursary, and David Walton Distinguished Doctoral Student Scholarship, which have generously supported me during my graduate studies at the University of Oxford.

Finally, I thank my family and friends for their constant encouragement and support throughout this research journey.

Contents

Chapter 1

Ambiguity Aversion, Portfolio Choice, and Life Expectancy*

Alistair Macaulay
University of Surrey

Chenchuan Shi
University of Oxford

Abstract

This paper studies how wealth and aging affect portfolio choices in a life-cycle model with ambiguity aversion. Ambiguity aversion implies that wealthier and older agents are endogenously more optimistic about risky asset returns, relative to poorer younger agents. As life expectancy grows, old agents become even more optimistic, while young agents become more pessimistic, amplifying the age gaps in portfolio composition, and implying further increases in intergenerational inequality. We find evidence for the mechanism in survey data on portfolios and subjective life expectancy. In a quantitative extension of the model, plausible life expectancy projections imply a 26% increase in the age-gradient of conditional risky asset shares between 2019 and 2100.

JEL codes: D84, E21, G11, J11

Keywords: Demographic Change, Ambiguity Aversion, Portfolio Choice, Inequality

*Macaulay (a.macaulay@surrey.ac.uk): School of Economics, University of Surrey. Shi (chenchuan.shi@nuffield.ox.ac.uk): Department of Economics & Nuffield College, University of Oxford. We thank Martin Ellison, Richard Foltyn, Tobias Kawalec, Xiaowen Lei, Carles Mañó-Cabello, and numerous conference participants for useful comments and suggestions. Declarations of interest: none. An earlier version of this paper was circulated under the title "Ambiguity Averse Portfolio Choices in an Aging Population."

1 Introduction

When faced with Knightian uncertainty, ambiguity averse agents over-weight the probability of the 'worst case' outcomes with low utility realizations (Gilboa and Schmeidler, 1989; Hansen and Sargent, 2008). These belief distortions differ according to the agent's environment, preferences, and characteristics: the worst case is very different, for example, for wealthy and poor households (Michelacci and Paciello, 2020, 2023), or between different policy regimes (Ferriere and Karantounias, 2019). As a consequence, when an economy goes through a structural shift, the belief distortions due to ambiguity aversion may change, with implications for a range of aggregate and distributional outcomes.

In this paper, we explore these effects for one of the key shifts taking place in developed economies today: population aging. In particular, we ask how increased lifespans affect ambiguity-averse beliefs about asset returns, and what that implies for portfolio choices and asset prices. This is likely to be important for policy in the coming decades, as savings behavior has been at the forefront of recent discussions of economic policy under demographic change (Goodhart and Pradhan, 2020; Auclert, Malmberg, Martenet and Rognlie, 2021; Kopecky and Taylor, 2022). Moreover, there is substantial empirical evidence that ambiguity aversion is a key determinant of portfolio choices and asset prices (Antoniou, Harris and Zhang, 2015; Dimmock, Kouwenberg, Mitchell and Peijnenburg, 2016; Collard, Mukerji, Sheppard and Tallon, 2018; Corgnet, Hernán-González and Kujal, 2020).

To analyze the interaction between aging, portfolio choice, and ambiguity aversion, we build an overlapping-generations model with ambiguity over risky asset returns. We find that increases in life expectancy cause young agents to distort their beliefs more strongly towards pessimistic outcomes, while older agents in contrast become more optimistic. Population aging therefore increases the concentration of equity holdings among older households, as their relatively more optimistic beliefs drive them to choose higher risky asset shares than the young. This in turn implies older generations become relatively wealthier, as they earn greater asset returns than younger agents. The key to this result is that increases in life expectancy have opposing effects on the marginal utility of income across age groups, leading to differing desires to insure against low-return outcomes.

This prediction is not present in models with aging that do not consider ambiguity aversion. To test it, we turn to survey data from the US. Consistent with the model, we find that the *age gradient* of equity shares is strongly increasing in subjective life expectancy. That is, among young households, an expectation of a longer lifespan is associated with less risky portfolios. Among older households, that pattern is reversed. This mechanism is important: a quantitative version of the model predicts that plausible longevity increases over the next 80 years will cause the age-profile of risky asset shares to steepen by 26%. This is a substantial shift, given that even the current mildly positive age gradient is already a major puzzle in the portfolio choice literature, which standard models struggle to replicate (see e.g. Ameriks and Zeldes, 2004; Chang, Hong and Karabarbounis, 2018). Our results therefore suggest that as the demographic transition continues it will become increasingly important to incorporate ambiguity aversion into household finance models.

As in Eggertsson, Mehrotra and Robbins (2019) and Malmendier, Pouzo and Vanasco (2020), we mostly focus on a simple case of the model with maximum three-period lifespans. This allows us to obtain analytic results, and inspect the mechanisms at work. We model ambiguity using so-called 'multiplier preferences', in which agents whose payoffs are more exposed to ambiguity distort their beliefs more strongly towards low-utility states (Hansen, Sargent and Tallarini, 1999). Although all investors view low returns as the worst case, a given drop in returns is much more damaging to some investors than others, and those who would be particularly badly affected do more to insure themselves against that possibility.[1]

We begin our analysis by considering an economy in which both the return on safe assets, and the distribution of returns on risky assets, are constant. We find that ambiguity aversion endogenously generates return expectations which are increasing in wealth and age. This is consistent with survey evidence on return expectations (Giglio, Maggiori, Stroebel and Utkus, 2021), experiments on biases in financial decision-making (Kovalchik, Camerer, Grether, Plott and Allman, 2005), and the observation that young and poor households typically hold less risky portfolios than those further up the age and wealth distribution (Chang et al., 2018; Catherine, 2022). Note that we match this last result despite the fact that, in the absence of ambiguity, the model reduces to a simple Merton (1969)-style model in which risky asset shares are constant for all agents.

Importantly, the effects of age on responses to ambiguity are driven by changes in *life expectancy*, rather than the number of years an agent has lived. As a consequence, reductions in mortality rates for older agents cause changes in the age profile of risky asset shares. Specifically, as mortality rates fall, young and old agents adjust their return expectations in opposite directions. The young become more pessimistic, distorting beliefs more strongly towards low risky asset returns, and investing less as a result - while simultaneously older agents become more optimistic, and invest more. As populations age around the world, this has consequences for the future of intergenerational inequality

[1] Several existing models of ambiguity aversion in portfolio choice instead assume all investors distort beliefs to the lowest return among a given set of possibilities, the range of which is typically specified as an exogenous aspect of preferences. This distinction is discussed further in the Related Literature section below.

and the composition of asset demand.

These endogenous changes in responses to ambiguity with wealth and life expectancy come about because of the interaction of two distinct channels. The first is the 'wealth channel', in which an agent who is saving more is more exposed in monetary terms to low asset returns. They have more 'skin in the game', and so have a greater desire to make their decisions robust to returns ambiguity, implying more pessimistic expectations.[2] The second is the 'marginal utility channel', in which an agent who expects to have a high marginal utility of consumption in the next period suffers a greater utility loss from a given fall in wealth. A poor agent may not lose much in monetary terms from a fall in returns, but due to the curvature of standard utility functions, they have a high marginal utility of consumption, and so of wealth. A small monetary loss can therefore have large utility consequences for these agents.

As life expectancy increases, younger agents save more, to fund consumption in their now-extended old age. They also expect to do the same throughout their middle age, implying lower consumption in the immediate future. With a diminishing marginal utility of consumption, this means that through both channels they become more exposed to shortfalls in asset returns, so they distort their beliefs more towards low returns to make their decisions more robust to ambiguity. As a result they invest less in risky assets, and instead allocate more of their savings to the risk-free asset.

For agents in middle age, the wealth channel operates in the same way. However, the marginal utility channel is reversed. When the probability of surviving into old age is low, they do not save much for those potential future periods. Conditional on survival, their old-age consumption is therefore low. As the mortality rate falls and life expectancy rises, they save more for their old age, which implies greater consumption in that period. As such, rising life expectancy reduces the expected marginal utility of consumption in old age, implying utility is *less* exposed to a given fall in wealth. Through the marginal utility channel, middle-aged agents therefore become more optimistic about asset returns, and increase their portfolio share in risky assets.

Which of these channels dominates is regulated by a simple condition: with CRRA preferences, the marginal utility channel dominates if and only if the elasticity of intertemporal substitution (EIS) is less than 1. In that case marginal utility changes sharply with (future) wealth, and so outweighs the wealth channel. Intuitively, this is related to the standard result that income effects of interest rate changes dominate substitution effects when the EIS is less than 1 (see Flynn, Schmidt and Toda, 2023, for an extended discussion). Substitution effects are small when marginal utility is highly convex, and this

[2]This echoes the literature arguing wealthier households have more incentive to process information about asset returns, for the same reason (Arrow, 1987; Lei, 2019; Macaulay, 2021).

is precisely when our marginal utility channel is large. Attempts to measure the EIS among households typically find values below 1 (Havránek, 2015), so we take this as our baseline.

A similar logic drives the effects of wealth on return expectations. An increase in wealth implies agents save more, making them more exposed to return fluctuations. At the same time, they expect to be wealthier in the future, which reduces their expected marginal utility of wealth. As with changes in life expectancy, the marginal utility channel dominates whenever the EIS is less than 1, in which case wealth reduces the extent to which agents distort return beliefs due to ambiguity aversion.

After characterizing these channels, we extend the model to allow for an endogenous equity premium. Initially, as life expectancy rises from a low level, the dominant force is the increasing optimism of the middle-aged agents. Aggregate demand for risk rises, and the equity premium falls. However, as life expectancy continues to grow, increases in middle-aged optimism slow down, and are eventually dominated by the greater pessimism of the young. Past a certain threshold, aggregate demand for risky assets begins to decline, and the equity premium rises as a result. This occurs because the effects of age on beliefs are smaller in the model for agents with more wealth. As mortality rates fall, young agents save more for their old age, middle-aged agents become wealthier, and so middle-aged agents become increasingly unresponsive to further increases in life expectancy. Interestingly, although the model is very stylized, this result is consistent with the U-shaped evolution of the equity premium observed across developed economies since 1950 (Kuvshinov and Zimmermann, 2020).

To test this mechanism, we use survey data to explore one of the most striking model predictions: that greater life expectancy is associated with a steeper age-gradient of risky asset shares (assuming an EIS below 1). This prediction is not present in other models in this literature, so offers a useful way to distinguish our model mechanism from others in the literature. In the Household Finance module of the US Survey of Consumer Expectations, households are asked about their portfolios and their subjective life expectancy. Using simple regressions of risky asset shares on age, interacted with subjective life expectancy, we find strong support for the model's prediction. The age-gradient of risky asset shares is close to zero for those with low subjective life expectancy, but is large and positive for those with high life expectancy.

Finally, we end with an illustration of how these effects might play out with plausible degrees of demographic change in the coming decades. We embed our model of ambiguity aversion into an otherwise-standard quantitative portfolio choice model, based on Gomes and Michaelides (2005), which includes risky age-dependent labor income, Epstein-Zin preferences, and equity market participation costs. We calibrate the model to data on mortality rates in the US in 2019, and to our regression results from the Survey of Consumer Expectations. The model produces risky portfolio shares with a similar level and age-gradient as those observed in the data, despite us not targeting these moments in the calibration. With this model, we simulate the effect of an increase in life expectancy, replacing the calibrated 2019 mortality rates with projected mortality rates for 2100. As in the analytical model, young agents become more pessimistic about asset returns relative to older agents, so the gap between the risky asset shares of old and young widens. Comparing agents at ages 80 and 35, the increase in life expectancy increases the gap in their risky asset shares by 26%. The simulated demographic changes to 2100 therefore have substantial consequences for portfolio decisions.

Related Literature: We principally contribute to the literature on ambiguity aversion, demographic change, and life-cycle portfolio choice in macroeconomics and finance.

Ambiguity aversion has been successful in providing theoretical explanations for a number of phenomena in macroeconomics and finance (see reviews in Ilut and Schneider (2023) and Epstein and Schneider (2010)). Our work is particularly related to models in which agents can endogenously adjust their response to ambiguity based on their own exposure to the variable(s) in question (Hansen et al., 1999; Cagetti, Hansen, Sargent and Williams, 2002; Bidder and Smith, 2012).[3] In particular, Michelacci and Paciello (2020, 2023) show that with ambiguity aversion wealthy and poor households hold systematically different expectations of aggregate variables. This explains several features of survey data on expectations, and influences macroeconomic dynamics. Our model extends these insights to portfolio choice, and shows that demographic changes therefore affect inequality and the equity premium.

Several of our results for how the belief distortions driven by ambiguity change with age and wealth depend on whether the elasticity of intertemporal substitution is greater or less than 1. In this we therefore add to the insights of Ferriere and Karantounias (2019) and Balter, Maenhout and Xing (2022), who show that the same condition determines the

[3]These examples, like us, use so-called 'multiplier preferences', in which agents distort beliefs towards a worst-case model, subject to a penalty of belief distortions which is linear in the relative entropy between the central and worst-case model used. An alternative approach that also allows endogenous continuous adjustment of belief distortions is the smooth ambiguity preferences of Klibanoff, Marinacci and Mukerji (2005, 2009), adapted to macroeconomic settings in Altug, Collard, Çakmaklı, Mukerji and Özsöylev (2020). Chen, Ju and Miao (2014) use this in a model of portfolio choice, but do not study changes over the life-cycle or with age. Multiple-priors models (as in Gilboa and Schmeidler, 1989) also feature endogenous belief distortions, but the optimal distortion is typically a corner solution. In Michelacci and Paciello (2020, 2023), for example, distortions to the perceived probability of certain shocks are either positive or negative, depending on a household's characteristics. Within those for whom a negative shock would harm utility, there are no variations in beliefs. In models like ours where the ambiguity is over asset returns, lower returns are always worse for utility, so multiple-priors models deliver a constant belief distortion for all agents.

outcome of models of optimal fiscal policy and momentum in asset return expectations respectively, once ambiguity aversion is present. Recent empirical evidence has tended to favor an EIS substantially below 1 (Havránek, 2015; Crump, Eusepi, Tambalotti and Topa, 2022), so we take this as the baseline case for our analysis.

In addition, our results are relevant for the literature on how demographic change will affect asset markets and inequality. The link between demography and asset markets was made prominent by a series of papers attempting to forecast what would happen as the large 'baby boomer' generation aged (the so-called 'asset market meltdown hypothesis', e.g. Poterba, 2001; Abel, 2001, 2003; Geanakoplos, Magill and Quinzii, 2004; DellaVigna and Pollet, 2007). This literature focuses largely on the consequences of changes in the relative size of older and younger investor populations. In contrast, the mechanisms we study concern how portfolio decisions *conditional on age* may change as expected lifespans grow.

This is important, because a variety of papers have studied demographic effects on aggregate asset demand by holding age profiles of asset holdings or savings rates fixed, and changing the proportions of households within each age group in line with past demographic data, or future projections (e.g. Mankiw and Weil, 1989; Mian, Straub and Sufi, 2021).[4] This approach only yields sufficient statistics for aggregate asset demand if household decision rules depend on that household's age, but are otherwise independent of aggregate demographic change (Auclert et al., 2021). We show that under ambiguity aversion, that is not the case, as changes in life expectancy affect decision rules.

Finally, we also relate to the large literature on portfolio choice over the life cycle (see Gomes, Haliassos and Ramadorai, 2021, for a review). Within this literature, a number of papers have proposed mechanisms to explain why older households typically invest the same or greater shares of their wealth in risky assets than young households. This pattern, while not present in standard life-cycle models (Cocco, Gomes and Maenhout, 2005; Gomes, 2020), can be generated by declines in labor market uncertainty as households age (Chang et al., 2018), or the cyclicality of return skewness (Catherine, 2022). Indeed, like us, Campanale (2011) and Peijnenburg (2018) offer explanations of the data based on ambiguity aversion.

We view the mechanism in this paper as complementary to these other forces. The key conceptual distinction between us and the existing literature is that in many of these previous papers, portfolio choices depend explicitly on the number of years an agent has lived to date. In contrast, the endogenous responses to ambiguity in our model imply

[4] An alternative approach analyses quantitative life-cycle models with rational expectations (e.g. Carvalho, Ferrero and Nechio, 2016; Kopecky and Taylor, 2022), which also abstract from the mechanisms we study.

that return expectations and portfolio decisions depend on the number of years an agent *expects* to live in the future.[5] In Peijnenburg (2018), for example, savers face Knightian uncertainty over a bounded interval of possible mean asset returns. With every period of life, they observe some returns data, and so can shrink that interval. Ambiguity-averse agents in that model set return expectations to the lower bound of the plausible interval, so the learning implies expected risky asset returns rise with age.[6] In our model, we instead consider a constant preference for robustness, and abstract from reductions in ambiguity through learning.[7] In this environment, we show that older households are typically less vulnerable to return shortfalls, and so choose to react less to their ambiguity. Ambiguity aversion therefore generates an upward-sloping age profile of risky asset shares, even if young poor households learn nothing about asset markets, as would be the case if they choose not to pay attention to them (as in e.g. Lei, 2019; Macaulay, 2021). Importantly, our mechanism also means that the age-profile of risky asset shares changes with life expectancy: among those with greater (subjective) life expectancy, older households are even more optimistic about asset returns relative to the young, and so the age gradient of risky asset shares is steeper. This accounts for the facts we document in survey data, which cannot be explained with other models in this literature.

The rest of the paper is organized as follows. Section 2 sets out the model environment with a general maximum lifespan. Section 3 characterizes the effects of ambiguity aversion and demographic change analytically in the special case where the maximum lifespan is 3 periods. Section 4 tests predictions of the model in survey data. Section 5 returns to the model, adds standard features of quantitative portfolio-choice models, and explores the implications for the age profile of risky asset shares both now and in the future. Section 6 concludes.

[5]Note that this implies the same mechanisms operate at the individual level for any increase in subjective survival probabilities, even if that is not reflected in objective reality (such deviations are studied in e.g. Grevenbrock, Groneck, Ludwig and Zimper (2021)). We abstract from this in this model to highlight the effects through returns ambiguity, though we make use of it in testing the model predictions in survey data.

[6]Similarly, in Chang et al. (2018) labor market uncertainty declines with age, enabling them to take more risk in their asset portfolios.

[7]A related model with multiplier preferences is Maenhout (2004). However, in that model the preference for robustness is normalized by wealth. While this normalization makes the model more tractable, it also assumes away many of the changes in belief distortions we study, and so abstracts from the mechanisms in our model.

2 Model

2.1 Environment

Demographics: Time is discrete. In each period t, a continuum of agents with measure 1 are born with age $j = 0$. An agent of age j survives to age $j+1$ in the next period with exogenous probability ϕ_j. We set $\phi_J = 0$, implying a maximum age of J. There is no population growth.

Preferences: An agent of age j chooses consumption and their portfolio allocation to maximize expected discounted lifetime utility:

$$U_{j,t} = E_{j,t} \sum_{k=j}^{J} \left[\beta^{k-j} \Phi_{j,k} \frac{c_{k,t+k-j}^{1-\gamma}}{1-\gamma} \right] \tag{1}$$

where β is the discount factor, $c_{j,t}$ is the consumption of an agent with age j in period t, and γ is relative risk aversion. $\Phi_{j,k}$ is the cumulative probability of surviving to age k, conditional on having survived to age j, defined as:

$$\Phi_{j,k} = \begin{cases} 1 & \text{for } k = j \\ \prod_{x=j}^{k-1} \phi_x & \text{for } k > j \end{cases} \tag{2}$$

Ambiguity aversion affects choices because it distorts the expectations operator $E_{j,t}$ away from the expectations calculated under the objective probability distribution of future outcomes. Ambiguity averse agents overweight the probabilities of future states with low utility and underweight states with high utility (Gilboa and Schmeidler, 1989; Hansen and Sargent, 2008).

Endowment and Savings: Agents are born with no financial assets. They receive a stream of deterministic labor income $\{y_{j,t}\}_j^J$.

There are two assets available for savings: a risk-free bond with gross interest rate R^f, and a risky asset with a gross return of R_t. The return on the risky asset is such that $\log(R_t)$ has an i.i.d. Normal distribution with mean $\tilde{\mu}$ and variance σ^2. For the results below, it will be helpful to define $\mu = \tilde{\mu} + \sigma^2/2$, where μ is the logarithm of $E_t R_{t+1}$.

Denote $s_{j,t}$ and $d_{j,t}$ as the amount of risky assets and risk-free bonds purchased by an agent of age j in period t. The budget constraint therefore reads:

$$c_{j,t} + s_{j,t} + d_{j,t} = R_t s_{j-1,t-1} + R^f d_{j-1,t-1} + y_{j,t} \tag{3}$$

Define financial wealth $x_{j,t}$ as the right hand side of equation (3):

$$x_{j,t} \equiv R_t s_{j-1,t-1} + R^f d_{j-1,t-1} + y_{j,t} \tag{4}$$

Furthermore, define human wealth $h_{j,t}$ as the present discounted value of future labor income:

$$h_{j,t} \equiv \sum_{k=j+1}^{J} \frac{y_{k,t+k-j}}{(R^f)^{k-j}} \tag{5}$$

Following Angeletos (2007), we define the agent's 'effective wealth' $w_{j,t}$ as the sum of financial and human wealth:

$$w_{j,t} \equiv x_{j,t} + h_{j,t} \tag{6}$$

Finally, we impose that $w_{J+1,T+1} \geq 0$, where T denotes the time period in which the agent will be age J. This prevents agents from borrowing when they know for certain they are in their final period. There are no bequests, so by assumption if an agent dies with positive asset holdings those assets are destroyed.

This setup is somewhat simpler than standard life-cycle portfolio choice models, such as Gomes and Michaelides (2005). In particular, we abstract from income risk, bequests, and risky asset participation costs. These simplifications allow us to obtain an analytic solution for portfolio choices. We view this as crucial for understanding the mechanisms behind the interactions between aging and ambiguity aversion.[8] We include these standard quantitative model features in Section 5, and find that our results from the analytic model are qualitatively unchanged.

2.2 Value Functions

Let $\alpha_{j,t} \in [0, 1]$ be the share of the agent's saving out of period-t effective wealth invested in risky assets:[9]

$$\alpha_{j,t} = \frac{s_{j,t}}{w_{j,t} - c_{j,t}} = \frac{s_{j,t}}{s_{j,t} + d_{j,t} + h_{j,t}} \tag{7}$$

Given this, we can express the budget constraint in terms of effective wealth only:

$$\begin{aligned} w_{j+1,t+1} &= R_{t+1} s_{j,t} + R^f d_{j,t} + y_{j+1,t+1} + h_{j+1,t+1} \\ &= (w_{j,t} - c_{j,t})[\alpha_{j,t} R_{t+1} + (1 - \alpha_{j,t}) R^f] \end{aligned}$$

[8] A further advantage of this framework is that, in the absence of ambiguity aversion, the model implies a constant portfolio share in risky assets for all ages (Merton, 1969). All changes we find in portfolio composition over an agent's life cycle must therefore come from ambiguity aversion.

[9] In all of the analysis here and in Section 3 we focus on the interior solution where these constraints on $\alpha_{j,t}$ do not bind. They will however become relevant in Section 5.

where we use $y_{j+1,t+1} = R^f h_{j,t} - h_{j+1,t+1}$, which follows from the definition of $h_{j,t}$.

The agent's utility maximization problem can now be written as:

$$V_j(w_{j,t}) = \max_{c_{j,t},\alpha_{j,t}} \left\{ \frac{c_{j,t}^{1-\gamma}}{1-\gamma} + \beta\phi_j E[V_{j+1}(w_{j+1,t+1})] \right\} \tag{8}$$

subject to:

$$w_{j+1,t+1} = (w_{j,t} - c_{j,t}) R^p_{j,t+1} \tag{9}$$

$$w_{J+1,T+1} \geq 0 \tag{10}$$

The initial condition is that $w_{0,t_0} = y_{0,t_0} + h_{0,t_0} = \sum_{k=0}^{J} \frac{y_{k,t_0+k}}{(R^f)^k}$, where $t_0 \equiv t - j$ denotes the time period in which the agent is born. $R^p_{j,t+1}$ is the agent's total return on their portfolio:

$$R^p_{j,t+1} \equiv R^f + (R_{t+1} - R^f)\alpha_{j,t} \tag{11}$$

Note that since $R^p_{j,t+1}$ is a weighted average over a constant R^f and a log-normal variable R_t, it is not itself log-normal. We follow Campbell (1993) and proceed with a log-linear approximation to the relationship between log portfolio returns and log individual-asset returns, taken about the point with zero excess returns:[10]

$$\log(R^p_{j,t+1}) \approx r^f + \alpha_{j,t}(r_{t+1} - r^f) + \frac{1}{2}\alpha_{j,t}(1-\alpha_{j,t})\sigma^2 \tag{12}$$

where r_{t+1} and r^f denote the log returns on the risky and safe assets respectively. With this approximation, the budget constraint (9) becomes:

$$w_{j+1,t+1} = (w_{j,t} - c_{j,t}) \exp\left[r^f + \alpha_{j,t}(r_{t+1} - r^f) + \frac{1}{2}\alpha_{j,t}(1-\alpha_{j,t})\sigma^2 \right] \tag{13}$$

Solution Without Ambiguity: If there is no ambiguity, the expectations operator $E_{j,t}$ coincides with expectations formed under the objective probability distribution of returns. Proposition 1 gives the optimal consumption and portfolio allocations in this case.

PROPOSITION **1**

Solving the household optimization (8) *subject to* (13) *and* (10)*, without ambiguity aver-*

[10] This approximation is exact in the limit of continuous time (Campbell and Viceira, 2002).

sion, implies a value function of the form:

$$V_j(w_{j,t}) = A_j \frac{w_{j,t}^{1-\gamma}}{1-\gamma} \tag{14}$$

where

$$A_j = \begin{cases} \left(\frac{1+b_{j+1}}{b_{j+1}}\right)^{\gamma} & \textit{for } j = 0, ..., J-1 \\ 1 & \textit{for } j = J \end{cases} \tag{15}$$

$$b_{j+1} = \left[\beta\phi_j A_{j+1}[1+(1-\gamma)r^f + \frac{1}{2}(1-\gamma)\frac{(\mu-r^f)^2}{\gamma\sigma^2}]\right]^{-\frac{1}{\gamma}} \tag{16}$$

Optimal consumption and portfolio choices are given by:

$$\alpha^*_{j,t} = \frac{\mu - r^f}{\gamma\sigma^2} \tag{17}$$

$$c^*_{j,t} = \begin{cases} \frac{b_{j+1}}{1+b_{j+1}} w_{j,t} & \textit{for } j < J \\ w_{j,t} & \textit{for } j = J \end{cases} \tag{18}$$

Proof. Appendix A □

As is well-known, this type of problem implies that the proportion of the agent's portfolio invested in risky assets is constant over time and age (Campbell and Viceira, 2002). To simplify notation, we remove the j, t subscripts and denote the optimal risky asset share in the absence of ambiguity as α^*. The path of optimal consumption depends on future survival probabilities, and therefore varies over the life cycle.

2.3 Ambiguity Aversion

We consider the case in which there is ambiguity over the mean of the return on the risky asset, as in Peijnenburg (2018). Formally, while equation (11) accurately reflects the return an agent would receive on a given portfolio invested in period t, the agent considers a set of models under which returns are distorted away from this by an amount $\nu_{j,t}$:

$$R^p_{j,t+1} = R^f + (R_{t+1} - R^f + \sigma_1\nu_{j,t})\alpha_{j,t} \tag{19}$$

where σ_1 is the standard deviation of R_{t+1}:

$$\sigma_1 = \exp(\mu)\left(\exp(\sigma^2) - 1\right)^{\frac{1}{2}} \tag{20}$$

To model the agent's aversion to ambiguity, we follow Hansen and Sargent (2008) and rewrite their dynamic programming problem as:

$$V_j^\theta(w_{j,t}) = \max_{c_{j,t},\alpha_{j,t}} \min_{\nu_{j,t}} \left\{ \frac{c_{j,t}^{1-\gamma}}{1-\gamma} + \beta\phi_j \left[\frac{1}{2\theta}\nu_{j,t}^2 + E_{j,t}[V_{j+1}^\theta(w_{j+1,t+1})] \right] \right\} \tag{21}$$

subject to $w_{J+1,T+1} \geq 0$, the budget constraint (9), and the distorted returns process (19).

That is, the agent makes consumption and portfolio decisions based on a distorted law of motion for their assets, in which returns on the risky asset are systematically biased towards models for the risky return which deliver low continuation values in their optimization problem. In this way they make choices which are robust to their uncertainty over the process for risky asset returns.

However, the agent does not entertain an infinite set of models. Rather, they choose the distortion in the returns process behind their consumption and portfolio choices so that it minimizes expected utility, *plus* a cost of $\frac{1}{2\theta}\nu_{j,t}^2$. Intuitively, the parameter θ controls the agent's preference for robustness: larger values of θ imply the agent entertains larger deviations from the true returns process in equation (11). Just as with the consumption and portfolio choices, the belief distortion $\nu_{j,t}$ is re-optimized every period, and the agent takes future belief distortion choices into account when making their decisions in period t. For a detailed discussion of this approach to modeling ambiguity aversion, see Hansen and Sargent (2008) and the survey in Ilut and Schneider (2023).

This formulation means that agents consider larger distortions if their value functions are more sensitive to model misspecification; in these cases the need for robustness is greater. Equation (21) shows that value functions differ by age directly through survival probabilities ϕ_j. In addition, value functions depend on wealth $w_{j,t}$, which may be correlated with age. Through both of these channels, the optimal distortions to beliefs about risky asset returns will vary across the life cycle. Intuitively, although all agents share the same preference for robustness (they have the same θ), agents of different ages have different levels of exposure to changes in the return on risky assets.

Optimal Belief Distortion: We begin by solving the inner minimization problem, in which the agent chooses how much to distort their return expectations towards the 'worst-case scenario'. In this, the following result is helpful.

Lemma 1

Taking the same log-linear approximation approach as in (12) *to the distorted returns in*

(19), *we can write*

$$E_{j,t}[V^\theta_{j+1}(w_{j+1,t+1})] \approx E_{j,t}[V^\theta_{j+1}(w^*_{j+1,t+1})] + \frac{A_{j+1}}{1-\gamma}(w_{j,t} - c_{j,t})^{1-\gamma}(1-\gamma)\sigma\alpha_{j,t}\nu_{j,t} \quad (22)$$

where

$$w^*_{j+1,t+1} = (w_{j,t} - c_{j,t})[R^f + \alpha_{j,t}(R_{t+1} - R^f)] \quad (23)$$

is the next-period wealth the agent would achieve under the central model without return distortions.

Proof. Appendix A □

Substituting this into the Bellman equation (21), it is then straightforward to obtain the first order condition for the inner minimization problem:

$$\nu_{j,t} = -\theta A_{j+1}(w_{j,t} - c_{j,t})^{1-\gamma}\sigma\alpha_{j,t} \quad (24)$$

Consumption and Portfolio Allocation: Substituting the optimal distortion (24) into equation (21), the household chooses consumption and the share of savings invested in risky assets to solve:

$$V^\theta_j(w_{j,t}) = \max_{c_{j,t},\alpha_{j,t}} \left\{ \frac{c^{1-\gamma}_{j,t}}{1-\gamma} + \beta\phi_j \left[\frac{1}{2}\theta A^2_{j+1}(w_{j,t} - c_{j,t})^{2-2\gamma}\sigma^2\alpha^2_{j,t} + E_{j,t}[V^\theta_{j+1}(w_{j+1,t+1})] \right] \right\} \quad (25)$$

Proposition 2 characterizes the solution.

Proposition 2
The value function takes the form:

$$V^\theta_j(w_{j,t}) = A_j \frac{w^{1-\gamma}_{j,t}}{1-\gamma} + \theta B_j \frac{w^{2(1-\gamma)}_{j,t}}{2(1-\gamma)} + O(\theta^2) \quad (26)$$

where B_j is an age-dependent combination of model parameters, defined in Appendix A.

In the approximate solution where we drop terms in θ^2, optimal portfolio choice and consumption are given by:

$$\alpha_{j,t} = \alpha^* + \theta\alpha^* w^{1-\gamma}_{j,t}\Omega_{\alpha j} \quad (27)$$
$$c_{j,t} = c^*_{j,t} + \theta w^{2-\gamma}_{j,t}\Omega_{cj} \quad (28)$$

where $\alpha^*, c^*_{j,t}$ are the solutions without ambiguity defined in Proposition 1, and $\Omega_{\alpha j}, \Omega_{cj}$ are functions of $b_{j+1}, A_{j+1}, B_{j+1}$, defined in Appendix A.

Proof. Appendix A □

With $\theta = 0$, we therefore return to the standard expected-utility solution (Proposition 1). With some ambiguity aversion ($\theta > 0$), however, both portfolio and consumption decisions shift away from this solution. The deviations from the rational-expectations solution are directly proportional to the degree of ambiguity aversion θ. Importantly, the effect of ambiguity aversion depends on both the agent's wealth and, through $\Omega_{\alpha j}$ and Ω_{cj}, their expected future lifespan. This occurs because agents of different ages are differentially exposed to the return on risky assets, and so opt for different levels of belief distortion in response to their ambiguity aversion. The resulting distortions to beliefs, consumption, and portfolio shares are generally nonlinear functions of wealth and demographics. To explore the mechanisms analytically, we therefore consider a case with a particularly simple demographic process.

3 Results with Three-Period Lifespans

We now restrict the model to $J = 2$. Agents therefore live for a maximum of three periods: at ages $j = 0, 1, 2$ we refer to them as young, middle aged, and old respectively. Furthermore, we assume that $\phi_0 = 1$, so all agents survive at least to middle age. In this simple context, population aging therefore only occurs through an increase in ϕ_1, the probability of surviving to old age.

3.1 Consumption and Portfolio Allocation: No Ambiguity Case

Consider an agent born in period t. To understand the effects of ambiguity aversion in this environment, it is helpful to first examine the forces that drive consumption and saving in the baseline model without ambiguity. In this case, the portfolio share in risky assets is constant, as in equation (17). Applying the remaining elements of Proposition 1, we obtain closed-form solutions for consumption in each period of the agent's life.

PROPOSITION **3**
An agent with $\phi_0 = 1, \phi_2 = 0$ and initial effective wealth $w_{0,t}$ chooses consumption when

young and middle-aged according to:

$$c_{0,t} = \frac{\tilde{b}^2}{\phi_1^{\frac{1}{\gamma}} + \tilde{b}(1+\tilde{b})} w_{0,t} \tag{29}$$

$$c_{1,t+1} = \frac{\tilde{b}}{\phi_1^{\frac{1}{\gamma}} + \tilde{b}(1+\tilde{b})} R^p_{0,t+1} w_{0,t} \tag{30}$$

where $\tilde{b}$ is a strictly positive combination of age-independent parameters:

$$\tilde{b} = \left[\beta \left(1 + (1-\gamma)r^f + \frac{1}{2}\frac{(1-\gamma)}{\gamma}(\alpha^*)^2\right)\right]^{-\frac{1}{\gamma}} \tag{31}$$

Conditional on surviving to old age, they then have:

$$c_{2,t+2} = \frac{\phi_1^{\frac{1}{\gamma}}}{\phi_1^{\frac{1}{\gamma}} + \tilde{b}(1+\tilde{b})} R^p_{0,t+1} R^p_{1,t+2} w_{0,t} \tag{32}$$

Proof. Appendix A □

As the probability of surviving to old age (ϕ_1) increases, the incentive to save for consumption in old age rises, so agents consume less when they are young and middle-aged ($c_{0,t}$ and $c_{1,t+1}$ decrease). Savings are therefore depleted less quickly through the life cycle if life expectancy is longer.[11] For agents that do survive to old age, a greater ϕ_1 implies higher consumption $c_{2,t+2}$, due to the extra savings built up earlier in life.

These patterns are displayed in Figure 1, which plots the paths of consumption and saving over the life cycle for three different values of ϕ_1. When the probability of surviving to old age is low, agents consume a lot in their youth and middle age. If they do survive to old age, they therefore experience a large consumption drop. With a greater survival probability, young and middle-aged consumption is lower, and the subsequent consumption fall in old age is lower.[12]

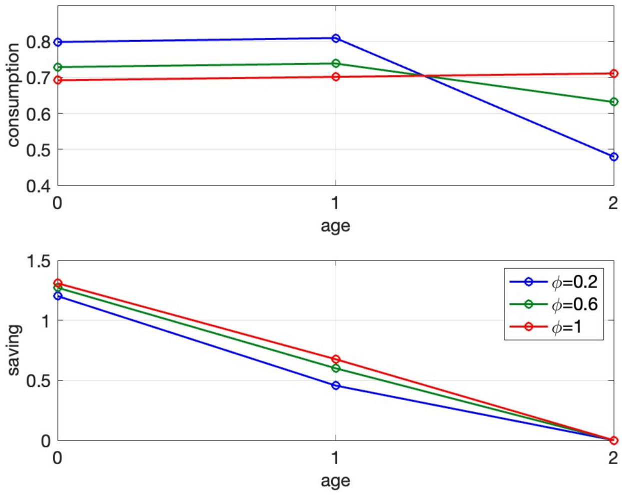
3.2 Ambiguity Aversion

We now add ambiguity aversion back into the model. The key element of this model is how the distortion in return expectations due to ambiguity aversion varies with age, wealth, and the probability of surviving to old age. These distortions are given in Proposition 4.

[11] This is consistent with Foltyn and Olsson (2024), who find that individuals with longer subjective life expectancies accumulate more wealth over their life cycle than those who expect to die earlier.

[12] With $\phi_1 = 1$ the consumption path is slightly increasing over time here due to precautionary saving. $c_{2,t+2} > c_{1,t+1}$ in the model whenever $\phi_1 > (\tilde{b}/R^p_{1,t+2})^\gamma$, which is close to 1 for most calibrations.

Figure 1: Consumption and saving paths with no ambiguity.



Note: Plots constructed using $J = 2$, $\mu = 0.06$, $r^f = 0.045$, $\sigma = 0.1$, $\gamma = 3$, $\beta = 0.99$, $\phi_1 = 0.2, 0.6, 1$, $w_{0,t} = 2$, and risky asset returns set to their expected level every period. This therefore abstracts from return shocks.

PROPOSITION **4**

The optimal distortion in beliefs about risky asset returns for an agent with $\phi_0 = 1, \phi_2 = 0$, wealth $w_{j,t}$, and age j is:

$$\nu_{0,t} = -\frac{\theta\sigma\alpha^*(\phi_1^{\frac{1}{\gamma}} + \tilde{b})}{\tilde{b}^\gamma(\phi_1^{\frac{1}{\gamma}} + \tilde{b} + \tilde{b}^2)^{1-\gamma}} w_{0,t}^{1-\gamma} \tag{33}$$

$$\nu_{1,t} = -\frac{\theta\sigma\alpha^*\phi_1^{\frac{1-\gamma}{\gamma}}}{(\phi_1^{\frac{1}{\gamma}} + \tilde{b})^{1-\gamma}} w_{1,t}^{1-\gamma} \tag{34}$$

$$\nu_{2,t} = 0 \tag{35}$$

Proof. Appendix A □

To understand the implications of these distortions, we first compare agents with the same wealth but different ages, to isolate the effects of age and survival probabilities. We then go on to analyse the interactions with varying wealth.

3.2.1 Age Effects

First, Proposition 4 implies that old agents ($j = 2$) do not distort their beliefs at all (35). This is because they save nothing, and so have no exposure to asset returns. There is no need for them to make their decisions robust to doubts about average asset returns. Similarly, note that if $\phi_1 = 0$ then a middle-aged agent sets $\nu_{1,t} = 0$, for the same reason: they will die for certain at the end of the period, so they do not save and are not exposed to ambiguity. In all other cases $\nu_{j,t} < 0$, so the agents distort their beliefs towards lower returns on the risky asset.

Corollary 1 shows a further equivalence between two other extreme special cases:

Corollary **1**
Let $\nu_{j,t}(\phi)$ be the distortion chosen by an agent of age j facing a survival probability of $\phi_1 = \phi$, as well as $\phi_0 = 1, \phi_2 = 0$. Then, if $w_0 = w_1$:

$$\nu_{0,t}(0) = \nu_{1,t}(1) \tag{36}$$

Proof. Appendix A □

That is, a young agent who will die for certain after middle age distorts beliefs in the same way as a middle-aged agent who will survive to old age for certain. In both cases, the agent knows they have one more period of consumption, and so behavior is the same for each. This highlights that life *expectancy*, rather than age, is the critical factor in how ambiguity aversion affects beliefs in this environment.

Second, Proposition 4 also implies that changes in ϕ_1 affect the belief distortions among young and middle-aged agents away from these edge cases.

Corollary **2**
For an agent with $\phi_0 = 1, \phi_2 = 0$, as ϕ_1 changes the optimal belief distortions of young agents, holding wealth constant, are such that:

$$\frac{\partial \nu_{0,t}}{\partial \phi_1} < 0 \tag{37}$$

Holding wealth constant, the distortions of middle-aged agents are such that:

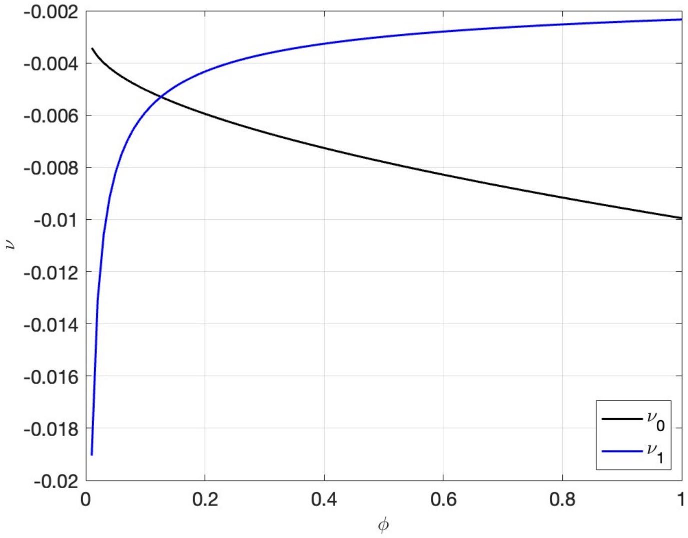
$$\frac{\partial \nu_{1,t}}{\partial \phi_1} \begin{cases} < 0 & \text{if } \gamma < 1 \\ = 0 & \text{if } \gamma = 1 \\ > 0 & \text{if } \gamma > 1 \end{cases} \tag{38}$$

Proof. Appendix A □

As the probability of surviving to old age rises, young households distort their beliefs more strongly, becoming more pessimistic about equity returns. If the EIS is greater than 1 ($\gamma < 1$), middle-aged households do the same. However, if the EIS is less than 1 ($\gamma > 1$), they decrease the magnitude of their distortion.

This divergence is at the heart of our mechanism: in the empirically plausible case with $\gamma > 1$, as life expectancy rises the young get more pessimistic about equity returns, while older agents get more optimistic. The effect is shown in Figure 2.

Figure 2: Belief distortions as ϕ_1 varies.



Note: Plots constructed using $J = 2$, $\theta = 0.045$, $\mu = 0.06$, $r^f = 0.045$, $\sigma = 0.1$, $\gamma = 3$, $\beta = 0.99$, $\phi_1 \in (0, 1]$, $w_{0,t} = w_{1,t} = 2$, *and risky asset returns set to their expected level every period. This therefore abstracts from the effect of return shocks. With this calibration,* $\sigma_1 \approx 0.01$, *so (using equation (19)) a distortion of* $\nu_{j,t} = -0.01$ *corresponds to a reduction in expected risky returns of approximately 10 basis points.*

At ϕ_1 close to 0, young households are less pessimistic than middle-aged households. As the survival probability grows, these positions reverse.

To understand the mechanisms driving the divergent responses to increasing longevity, we return to the first-order condition for the inner minimization in equation (21). The agent chooses the degree to which they distort their return expectations by balancing the marginal damage to expected continuation values with the marginal penalty to considering a larger distortion:

$$\frac{\partial}{\partial \nu_{j,t}} \left\{ \frac{1}{2\theta} \nu_{j,t}^2 + E_{j,t}[V_{j+1}^{\theta}(w_{j+1,t+1})] \right\} = 0 \tag{39}$$

Equation (24) is then simply the result of combining this with Lemma 1 and rearranging. However, we can alternatively write this condition as:

$$\begin{aligned} \nu_{j,t} &= -\theta \frac{\partial E_{j,t}[V_{j+1}^{\theta}(w_{j+1,t+1})]}{\partial \nu_{j,t}} && (40) \\ &\approx -\theta\sigma\alpha^* \underbrace{(w_{j,t} - c_{j,t})}_{\text{Wealth Channel}} \cdot \underbrace{\frac{\partial E_{j,t}[V_{j+1}^{\theta}(w_{j+1,t+1})]}{\partial w_{j+1,t+1}}}_{\text{Marginal Utility Channel}} && (41) \end{aligned}$$

where the second line uses the same approximation as in Lemma 1.

The distortion is set proportional to the sensitivity of expected continuation values to asset returns. Intuitively, the more exposed the agent is to changes in risky asset returns, the more they wish to make their decisions robust to ambiguity over those returns. That sensitivity can be broken down into two channels: the wealth channel, and the marginal utility channel.

The wealth channel operates because when an agent saves more for the next period, their next-period wealth is more strongly affected by asset returns. In other words, they have more skin in the game. As discussed in Section 3.1, when ϕ_1 increases both young and middle-age agents increase their saving.[13] For both young and middle-age agents, this channel therefore implies greater belief distortions when ϕ_1 rises.

The marginal utility channel operates because a given decrease in asset returns will have a larger effect on utility for an agent with a large marginal utility of wealth in the following period. Through a standard envelope theorem, the marginal utility of wealth in period $t+1$ is equal to the marginal utility of consumption in $t+1$. Since our model features a diminishing marginal utility of consumption, this channel will be more powerful when next-period consumption is expected to be low.

This channel is what drives the divergence in beliefs across cohorts. Recall from Section 3.1 that, as ϕ_1 increases, the consumption of middle-aged agents falls, while the consumption of old agents rises. Agents who are currently young therefore expect to have a greater marginal utility of wealth in the following period, when they will be middle-aged. An increase in ϕ_1 makes them more sensitive to changes in wealth, increasing the strength of the marginal utility channel. In contrast, current middle-aged agents expect to have more wealth in future, and so a lower marginal utility, implying a smaller marginal utility channel.

For a young agent, both the wealth and marginal utility channels therefore imply that they become more pessimistic when ϕ_1 rises. For a middle-aged agent, the channels act in opposite directions. To see which dominates, note that for a middle-aged agent we obtain:[14]

$$\frac{\partial E_{1,t}[V_2^\theta(w_{2,t+1})]}{\partial w_{2,t+1}} \propto (w_{1,t} - c_{1,t})^{-\gamma} \tag{42}$$

[13]The paths of consumption and saving remain qualitatively unchanged with the introduction of ambiguity aversion, as we typically consider small values of θ. Sufficient conditions for this result, and a numerical example, are provided in Appendix B.

[14]This follows from Proposition 2, and the fact that $A_2 = 1$.

Substituting this into equation (41) implies:

$$\nu_{1,t} \propto -\theta\sigma\alpha^{*}(w_{1,t} - c_{1,t})^{1-\gamma} \tag{43}$$

For a middle-aged agent, an increase in ϕ_1 implies a rise in $w_{j,t} - c_{j,t}$. With $\gamma < 1$, the wealth channel dominates, and middle-aged agents therefore increase the magnitude of their belief distortions when survival probabilities rise ($\nu_{j,t}$ becomes more negative). In the empirically plausible case with $\gamma > 1$, however, the marginal utility channel dominates, and middle-age agents become more optimistic about returns. With log utility ($\gamma = 1$) the effects cancel out and middle-aged agents do not adjust $\nu_{1,t}$ with ϕ_1.

Importantly, these channels depend on how consumption and saving *change* in response to an increase in longevity, not on their level. This explains why the mechanisms continue to operate in richer models with more realistic income processes, such as the one in Section 5. Quantitative models of this kind share the prediction that agents save more for old age, and consume more if they survive, when survival rates increase (Gomes, 2020).

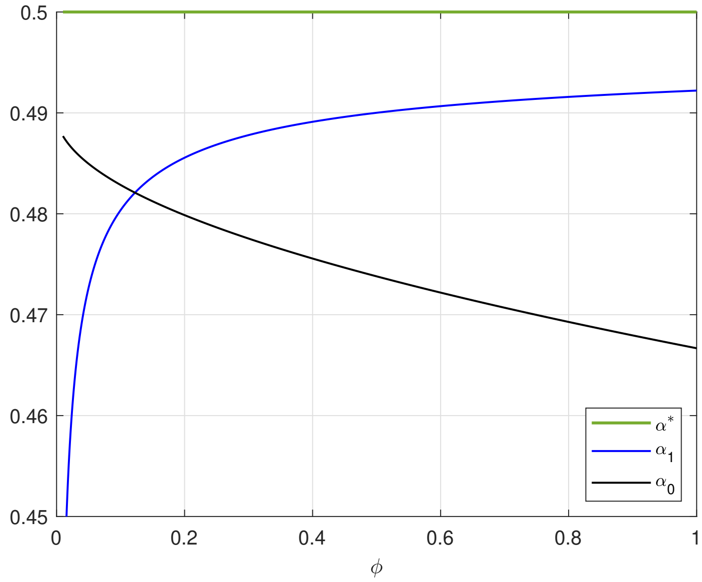
Portfolio Allocation: Figure 3 plots the share of agent portfolios invested in the risky asset ($\alpha_{j,t}$) as ϕ_1 changes, for the same parameters as Figure 2. Both young and middle-aged agents allocate lower shares of their wealth to risky assets than they would in the absence of ambiguity, as in other models in the literature (Garlappi, Uppal and Wang, 2007; Campanale, 2011) and consistent with empirical evidence (Dimmock et al., 2016). This is a direct consequence of Proposition 4: the belief distortions due to ambiguity aversion imply the agent acts as if the risky asset has a lower expected return than its true mean, and so invests less in that asset than they would in the absence of ambiguity. For $\phi_1 > 0.2$ in this calibration, young agents distort their return beliefs more in response to ambiguity than middle-aged agents, so their risky asset share is correspondingly lower. We find the same qualitative pattern in the quantitative illustration in Section 5.

The changes in risky asset shares as the population ages (ϕ_1 increases) follow from Corollary 2. As ϕ_1 increases along the x-axis, middle-aged agents reduce the distortions in their beliefs, increasing their risky asset share towards the benchmark share without ambiguity (α^*).[15] In contrast, young households increase their distortions, and move further from this benchmark level.

Note that the models in Campanale (2011) and Peijnenburg (2018) also feature risky asset shares that increase with age ($\alpha_{1,t} > \alpha_{0,t}$). However, the mechanisms in those papers are different from ours: there agents learn over time from observed realizations of

[15] Equation (34) highlights that even with $\phi_1 = 1$ the belief distortion is strictly negative, so $\alpha_{1,t}$ always remains strictly below α^*.

Figure 3: Risky asset shares without ambiguity (α^*), and with ambiguity for young (α_0) and middle-aged (α_1) agents, as ϕ_1 varies.



Note: Plots constructed using $J = 2$, $\theta = 0.045$, $\mu = 0.06$, $r^f = 0.045$, $\sigma = 0.1$, $\gamma = 3$, $\beta = 0.99$, $\phi_1 \in (0, 1]$, $w_{0,t} = w_{1,t} = 2$, *and risky asset returns set to their expected level every period. This therefore abstracts from the effect of return shocks.*

asset returns, gradually reducing the set of models they are willing to consider. In our framework, that would entail a fall in θ as agents progress from young to middle-aged, independently of survival probabilities. In contrast, we keep θ constant, but allow the optimal distortion to vary with agent exposure to asset returns. The effect of survival probabilities on the age-profile of asset shares, through the wealth and marginal utility channels, is therefore unique to our mechanism.

3.2.2 Wealth Effects

The first-order condition for belief distortions (41) highlights that, just as with age, the effects of an increase in wealth depend on the wealth and marginal utility channels.

With an increase in wealth, these channels act in opposite directions. The wealth channel implies larger belief distortions for wealthier households, as they save more, so have more exposure to asset returns. The marginal utility channel implies the opposite: wealthier households have smaller belief distortions, because their continuation values are less sensitive to marginal changes in future wealth. As with the effect of survival probabilities on middle-aged agents, which channel dominates depends on whether the EIS is greater or less than 1.

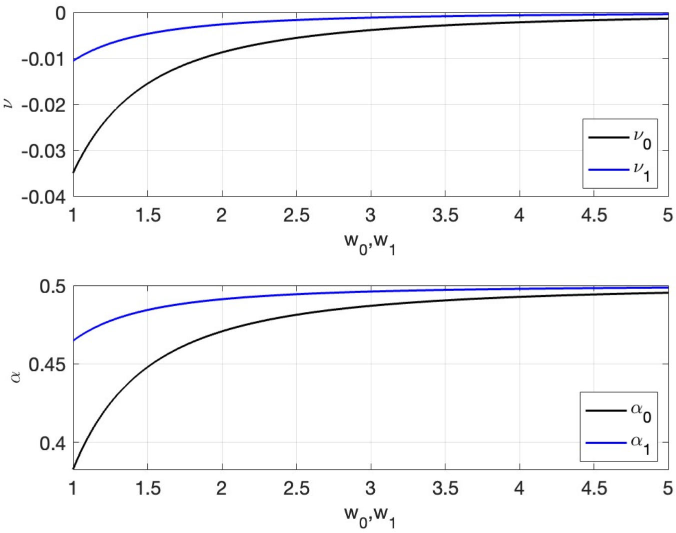
Corollary **3**

For agents with $\phi_0 = 1, \phi_2 = 0$, $w_{j,t}$ affects optimal belief distortions such that for $j = 0, 1$:

$$\frac{\partial \nu_{j,t}}{\partial w_{j,t}} \begin{cases} < 0 & \textit{if } \gamma < 1 \\ = 0 & \textit{if } \gamma = 1 \\ > 0 & \textit{if } \gamma > 1 \end{cases} \tag{44}$$

Proof. Appendix A □

Figure 4: Belief distortions and risky asset shares vary with wealth.



Note: Plots constructed using $J = 2$, $\theta = 0.045$, $\mu = 0.06$, $r^f = 0.045$, $\sigma = 0.1$, $\gamma = 3$, $\beta = 0.99$, $\phi_1 = 0.7$, $w_{0,t}$ and $w_{1,t}$ vary from 1 to 5, and risky asset returns set to their expected level every period. This therefore abstracts from the effect of return shocks.

Under our preferred calibrations ($\gamma > 1$), being wealthier causes agents to become more optimistic about asset returns. As a result, wealthier agents invest a greater share of their wealth in the risky asset. These patterns are shown in Figure 4. Although this analytic model does not feature non-participation, the direction of this effect is consistent with evidence in Briggs, Cesarini, Lindqvist and Östling (2020) that exogenous increases in wealth increase the probability that households invest in equities.

Interactions with Age Effects: As well as directly affecting belief distortions as in Corollary 3, an agent's wealth can affect the strength of the age effects on beliefs studied in Section 3.2.1. In our preferred parameter region of $\gamma > 1$, when wealth is higher, age effects are smaller in magnitude, for agents of all ages.

Corollary 4

For agents with $\phi_0 = 1, \phi_2 = 0$, the age effects on optimal belief distortions change with wealth such that for $j = 0, 1$:

$$sign\left\{\frac{\partial}{\partial w_{j,t}}\left(\frac{\partial \nu_{j,t}}{\partial \phi_1}\right)\right\} = \begin{cases} sign\left\{\dfrac{\partial \nu_{j,t}}{\partial \phi_1}\right\} & if\, \gamma < 1 \\ 0 & if\, \gamma = 1 \\ -sign\left\{\dfrac{\partial \nu_{j,t}}{\partial \phi_1}\right\} & if\, \gamma > 1 \end{cases} \tag{45}$$

This implies that if $\gamma > 1$, the effect of ϕ_1 on optimal belief distortions decreases in magnitude as $w_{j,t}$ increases.

Proof. Appendix A □

Intuitively, at high levels of wealth, future marginal utility is less sensitive to changes in returns, and so the marginal utility channel is weakened. When $\gamma > 1$, the marginal utility channel is the dominant channel determining how belief distortions $\nu_{j,t}$ change with ϕ_1 at all ages. Weakening that channel therefore weakens the effects of ϕ_1 on beliefs.

Throughout this analysis we maintain the simplifying assumption that any wealth left un-consumed by agents who die before reaching their maximum lifespan is destroyed. One consequence of Corollary 4 is that this assumption, if anything, implies we will understate the magnitude of the mechanisms we study.

Specifically, a common alternative assumption in this literature is that any such un-consumed wealth is left as an 'accidental bequest', and so is distributed between surviving agents in the following period (e.g. Gagnon, Johannsen and López-Salido, 2021). If we assumed this, then a rise in the survival probability ϕ_1 would imply fewer agents die before old age, which in turn means fewer agents leave accidental bequests, which reduces the wealth of all agents.[16] With that lower wealth, Corollary 4 implies the direct effects of ϕ_1 on beliefs become stronger. The differences between young and middle-aged agents widen. To keep the exposition of the main channels as clear as possible, we continue to abstract from this endowment effect in the remainder of the paper.

[16]There would also be an offsetting effect, that greater ϕ_1 implies agents save more for old age, so conditional on dying before old age an agent leaves a larger accidental bequest. Since old-age consumption (and thus wealth) are concave in ϕ_1 for our preferred parameter range of $\gamma > 1$ (32), this effect is dominated by the simple channel of fewer early deaths as long as ϕ_1 is not too small. In the calibration used for the figures in this section, accidental bequests decline with ϕ_1 for all $\phi_1 > 0.032$.

3.3 Intergenerational Inequality

We now use the results developed above to analyze the impacts of increased longevity through ambiguity aversion. The first implication we study concerns how wealth evolves over an agent's life-cycle. If realizations of the risky asset return are close to its mean, this is also informative about intergenerational wealth inequality, as there are no other shocks that change from one cohort to the next.

Using equation (13), we can express the ratio between wealth at ages $j+1$ and j as:

$$\frac{w_{j+1,t+1}}{w_{j,t}} = \left(1 - \frac{c_{j,t}}{w_{j,t}}\right) \exp\left(r^f + \alpha_{j,t}(r_{t+1} - r^f) + \frac{1}{2}\sigma^2 \alpha_{j,t}(1-\alpha_{j,t})\right) \tag{46}$$

Expanding out $c_{j,t}$ and $\alpha_{j,t}$ using Proposition 2, this becomes:

$$\begin{aligned}\frac{w_{j+1,t+1}}{w_{j,t}} = \left(\frac{1}{1+b_{j+1}} - \theta w_{j,t}^{1-\gamma}\Omega_{c,j}\right) \exp\Bigg(& r^f + \alpha^*(r_{t+1} - r^f) + \frac{1}{2}\sigma^2\alpha^*(1-\alpha^*) \\ &+ \theta\alpha^* w_{j,t}^{1-\gamma}\Omega_{\alpha,j}\left[r_{t+1} - r^f + \frac{1}{2}\sigma^2(1 - 2\alpha^* - \theta\alpha^* w_{j,t}^{1-\gamma}\Omega_{\alpha,j})\right]\Bigg)\end{aligned} \tag{47}$$

Finally, using the definition of b_{j+1} for $j = \{0,1\}$ (Appendix A) note that:

$$\frac{1}{1+b_1} = \frac{\phi_1^{\frac{1}{\gamma}} + \tilde{b}}{\phi_1^{\frac{1}{\gamma}} + 2\tilde{b}} \tag{48}$$

$$\frac{1}{1+b_2} = \frac{\phi_1^{\frac{1}{\gamma}}}{\phi_1^{\frac{1}{\gamma}} + \tilde{b}} \tag{49}$$

both of which are strictly increasing in ϕ_1.

In the absence of ambiguity ($\theta = 0$), only the first term in equation (47) changes with ϕ_1. As the survival probability rises, agents save more for the future, so middle-aged agents become wealthier relative to young agents, and similarly old agents become wealthier relative to middle-aged agents.

Ambiguity adds two further channels to this change in wealth across the life-cycle. First, a rise in ϕ_1 affects agent portfolio choices, and so affects average returns. This generates the terms in square brackets in equation (47). In Section 3.2.1 we showed that rising ϕ_1 has opposing effects on the risky asset shares of young and middle-aged agents in the empirically plausible case of $\gamma > 1$. Young agents reduce $\alpha_{0,t}$, which reduces the wealth of the middle-aged relative to the young. Middle-aged agents increase $\alpha_{1,t+1}$, increasing the relative wealth of the old. Through this channel, an aging population leads to a greater wealth concentration among older households.

Second, ambiguity also affects the amount saved by each agent, through the term $-\theta w_{j,t}^{1-\gamma}\Omega_{c,j}$. For both young and middle-aged agents, $\Omega_{c,j}$ varies with ϕ_1, and for middle-aged agents so does $w_{j,t}$. In Appendix A (equation (140)) we show that this term is potentially non-monotonic in ϕ_1, so it has an ambiguous effect on wealth inequality. However, in the simple calibration used throughout this section this effect is negligible relative to the other channels.[17]

These effects concern how changes in life expectancy affect the relative wealth between generations. In this simple model, wealth typically declines with age, because we lack many of the features commonly added to quantitative life-cycle models.[18] The mechanisms identified in this section imply that a rise in ϕ_1 causes a greater concentration of wealth among older generations relative to this baseline. When we extend the model for our quantitative exercises in Section 5, the life-cycle of wealth in the model matches the hump-shaped pattern observed in the data, but the mechanisms identified here continue to operate.

3.4 Aggregation

Next, we study how aging affects the composition of aggregate asset demand.

Recall that each period a new cohort of agents is born with measure 1, so there are 3 overlapping generations alive in each period. All agents within a cohort are identical. The aggregate demand for safe and risky assets is therefore given by:

$$AD_t(safe) = (1-\alpha_{0,t})(w_{0,t}-c_{0,t}) + (1-\alpha_{1,t})(w_{1,t}-c_{1,t}) \tag{50}$$

$$AD_t(risky) = \alpha_{0,t}(w_{0,t}-c_{0,t}) + \alpha_{1,t}(w_{1,t}-c_{1,t}) \tag{51}$$

where we have used the result that old agents do not save in either asset, and that all agents survive to middle age. Note that this implies the composition effects of aging, as studied in e.g. Auclert et al. (2021), are absent here: the age-composition of *asset market participants* is constant as ϕ_1 rises. This simplification allows us to focus on the novel channels introduced by ambiguity aversion here.

We showed previously that if there is no ambiguity aversion ($\theta = 0$), risky asset shares α_j are constant, and both young and middle-aged agents cut consumption when ϕ_1 rises

[17]Specifically, with the calibration used in Figure 2, an increase in ϕ_1 from 0.2 to 0.8 implies that $w_{2,t+2}/w_{1,t+1}$ and $w_{1,t+1}/w_{0,t}$ increase by 31.02% and 7.48% respectively. The change in consumption due to ambiguity accounts for 0.014% and 0.019% of those changes.

[18]We use the word 'typically' here because if there is a very large realization of the risky asset return, it is possible for wealth to increase from age j to $j+1$.

(Proposition 3). As a result, the aggregate demand for both types of asset rises, as longer life expectancy encourages greater saving for old age.

In the case with ambiguity aversion, the deviation of $\alpha_{j,t}$ and $c_{j,t}$ from the no-ambiguity benchmark is proportional to the degree of ambiguity aversion θ (Proposition 2). For small θ, we therefore maintain the result that $AD_t(safe)$ and $AD_t(risky)$ rise with ϕ_1, as in the no-ambiguity benchmark.

Ambiguity aversion does, however, affect the speed at which each aggregate asset demand rises, which therefore affects the *composition* of asset demand as the population ages. This is displayed in equation (52), which gives the population analogue of the individual-level risky share $\alpha_{j,t}$.

$$\frac{AD_t(risky)}{AD_t(safe) + AD_t(risky)} = \frac{\alpha_{0,t}(w_{0,t} - c_{0,t}) + \alpha_{1,t}(w_{1,t} - c_{1,t})}{w_{0,t} - c_{0,t} + w_{1,t} - c_{1,t}} \tag{52}$$

Substituting out for the individual risky asset shares $\alpha_{j,t}$ using equation (27), this becomes:

$$\frac{AD_t(risky)}{AD_t(safe) + AD_t(risky)} = \alpha^* + \theta\alpha^* \left(\frac{\Omega_{\alpha,0} w_{0,t}^{1-\gamma}(w_{0,t} - c_{0,t}) + \Omega_{\alpha,1} w_{1,t}^{1-\gamma}(w_{1,t} - c_{1,t})}{w_{0,t} - c_{0,t} + w_{1,t} - c_{1,t}} \right) \tag{53}$$

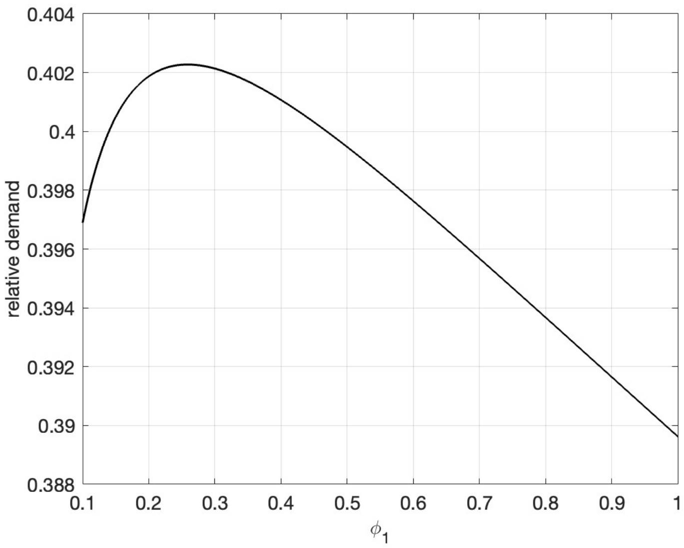
In the absence of ambiguity, the relative demand for risky assets is a constant (α^*). However, with ambiguity ($\theta > 0$), demand for safe and risky assets are no longer in fixed proportions, and the relative demand for each asset changes with the survival probability.

These relative demand changes are shown in Figure 5. The no-ambiguity relative demand is constant at α^*, which with this calibration is equal to 0.5. As in e.g. Garlappi et al. (2007), ambiguity aversion reduces the relative demand for risky assets below this level. The contribution of our model is that we can ask how that relative demand changes with survival rates. As ϕ_1 rises, the demand for risky assets relative to safe assets follows a hump-shape: it rises, reaches a peak, then falls.

This hump-shape in relative risky asset demand occurs because young and middle-aged agents shift their belief distortions in different directions as ϕ_1 increases, as shown in Section 3.2.1. As the survival probability rises, the young become more pessimistic about asset returns, while the middle-aged become more optimistic (Corollary 2). As a result, young agents decrease their relative demand for risky assets, while middle-aged agents increase their relative demand.

When the probability of surviving to old age is low, young households save only a small fraction of their endowment (Proposition 3, extended to the ambiguity case in

Figure 5: Relative asset demand $\frac{AD_t(risky)}{AD_t(safe)+AD_t(risky)}$ varies with ϕ_1.



Note: Plots constructed using $J = 2$, $\mu = 0.06$, $r^f = 0.045$, $\sigma = 0.1$, $\theta = 0.045$, $\gamma = 3$, $\beta = 0.99$, $\phi_1 \in (0, 1]$, $w_{0,t} = 1$, and risky asset returns set to their expected level every period. This therefore abstracts from the effect of return shocks.

Appendix B). As a result, when they become middle-aged, they have little wealth: $w_{1,t}$ is low relative to $w_{0,t}$. Corollary 4 then implies that any small increase in ϕ_1 has a stronger effect on the beliefs of middle-aged than young agents. Initially, the middle-aged agents react most strongly to ϕ_1, and relative risky asset demand rises.

However, as ϕ_1 rises, young agents increase their saving, and these mechanisms work in reverse. Wealth $w_{1,t}$ rises, and so age effects become weaker for middle-aged agents, ultimately becoming smaller than the effects on young agents. At high ϕ_1, further aging of the population therefore implies a fall in relative risky asset demand.

Finally, note that relative risky asset demand is also affected by a composition channel. As ϕ_1 rises, $w_{1,t}$ rises, which means that middle-aged agents account for a greater share of aggregate saving. Since for most values of ϕ_1 middle-aged agents are more optimistic than young agents (Figure 2), this also causes the aggregate relative risky asset demand to rise with ϕ_1, shifting the peak in Figure 5 to the right.

3.5 Endogenous Equity Premium

So far, we have considered a small open economy aging alone. In that case, the variation in relative demand for safe and risky assets shown in Figure 5 does not affect returns or asset prices. However, in a closed economy, or indeed in a world where all countries are aging simultaneously, this will no longer be true.

We therefore extend the model here, and instead assume that the relative supply of safe and risky assets is fixed. This allows us to study the effects of aging on the equity premium $\mu - r^f$, as this must adjust to ensure that asset markets clear. The equity premium is particularly of interest because it controls how much heterogeneity there is between the wealth accumulation of agents with different beliefs. It is therefore central to how our mechanisms will affect intergenerational inequality.

Specifically, let $S_t(safe)$ and $S_t(risky)$ denote the supply of safe and risky assets in period t. The relative supply of risky assets is assumed to be fixed at $\bar{S}$:

$$\frac{S_t(risky)}{S_t(safe) + S_t(risky)} = \bar{S} \tag{54}$$

For asset markets to clear, we therefore require:[19]

$$\frac{AD_t(risky)}{AD_t(safe) + AD_t(risky)} = \bar{S} \tag{55}$$

This particular assumption on asset supply is useful here, because it implies that if there is no ambiguity aversion, the solution is trivial. From equation (53), if $\theta = 0$ then the relative demand for risky assets is constant at α^*, which itself is directly proportional to the equity premium (equation (17)). In this case, the equilibrium equity premium is therefore a constant, unaffected by changes in survival probabilities:

$$(\mu - r^f | \theta = 0) = \gamma \sigma^2 \bar{S} \tag{56}$$

As a result, any dependence of the equity premium on ϕ_1 must come through ambiguity aversion. In this way, our equity premium analysis is similar in spirit to our analysis of the small open economy above, in which individual portfolio choices are independent of ϕ_1 unless there is some ambiguity over risky returns.

To analyze the equity premium in the case with ambiguity, it is useful to first note that relative aggregate demand increases monotonically in the equity premium.

LEMMA **2**

In the model with $\phi_0 = 1, \phi_2 = 0$, for any $\theta < \theta^*$:

$$\frac{\partial}{\partial \mu} \left(\frac{AD_t(risky)}{AD_t(safe) + AD_t(risky)} \right) > 0 \tag{57}$$

[19]This condition would still be necessary, though not sufficient, for asset market clearing in a model with fixed supplies of both assets individually.

$$\frac{\partial}{\partial r^f}\left(\frac{AD_t(risky)}{AD_t(safe)+AD_t(risky)}\right)<0 \tag{58}$$

where $\theta^* > 0$ is a threshold defined in Appendix A.

Proof. Appendix A $\square$

This is intuitive: as in the case without ambiguity, if the equity premium rises, then the expected return on risky assets rises relative to safe assets, rendering them more attractive to investors. Whether the equity premium rises because μ rises or r^f falls, the relative demand for risky assets therefore rises.

From this we arrive at two implications. First, Proposition 4 implies that for all values of ϕ_1, young and middle-aged agents distort their beliefs towards (weakly) lower risky asset returns. This is why ambiguity reduces the relative demand for risky assets, as shown in Figure 5. To offset this and ensure market clearing, the equity premium must therefore be higher than if there was no ambiguity. Ambiguity aversion can therefore help explain the equity premium puzzle, as in other models in this literature (e.g. Dimmock et al., 2016).

Second, the analysis in the previous sections highlights that individual and aggregate portfolio choices change with ϕ_1, implying that the equity premium will change as survival probabilities rise. Specifically, the equity premium is U-shaped in ϕ_1.

The intuition for this result follows directly from the discussion in Section 3.4. As ϕ_1 rises from a low level, then the relative aggregate demand for risky assets rises, as middle-aged agents become more optimistic about risky returns. This pushes the equity premium down, to clear asset markets. As ϕ_1 continues to rise, the relative aggregate demand for risky assets begins to fall, as increasing pessimism from young agents dominates the optimism from the middle-aged. That in turn implies the equity premium rises.

Interestingly, although the model is extremely simple, this is consistent with qualitative patterns in equity premia in the last 75 years. Since 1950, developed economies have experienced substantial rises in life expectancy. Over the same period, Kuvshinov and Zimmermann (2020) document that equity premia in developed economies have followed a U-shape, first falling, and then rising again after 1990. However, note that in Appendix D.7 we find that in our quantitative model the effects of life expectancy on equity premia are rather small, so the channel discussed here is unlikely to explain all of the historical equity premium experience.

4 Testing the Mechanism

The key novel prediction of our model is that *life expectancy*, not just the current age, affects beliefs and portfolio decisions. In this section, we test this prediction using survey data from the US. We find evidence in favor of the mechanisms outlined in Section 3: for young households, a longer subjective life expectancy is associated with smaller portfolio shares invested in equities. For older households, that relationship is reversed. This qualitative pattern is produced by our model for $\gamma > 1$, as shown in Figure 3.

4.1 Data

We use the Household Finance Module included in the August waves of the Federal Reserve Bank of New York Survey of Consumer Expectations (SCE) between 2014-2019.[20] In each of the six waves, approximately 1,100 households are asked detailed questions about their portfolio choices and, importantly for our purposes, their subjective life expectancy. The survey also collects rich demographic information, including age, gender, race, and education. The sample is designed to be representative of the US population.

Household Portfolios: The first question we use asks about the share of the respondent's portfolio invested in equities:

> "What proportion of the money in your (and your spouse's/partner's) saving and investment accounts (excluding funds in retirement accounts) is invested in stocks/stock mutual funds?"

This question is not conditional on asset market participation: people who do not invest in equities give an answer of 0.

Note that this question is similar, but not quite identical, to $\alpha_{j,t}$ in the simple model. The question elicits the share of financial wealth invested in stocks, and the model concept is the share of total (effective) wealth invested. We focus on the untransformed data in our main analysis for clarity. In Appendix C, we show that our results are robust to transforming this variable to measure the fraction of total wealth invested in stocks.

As well as portfolio shares, we also use the information recorded in the survey on household income. We use the many questions on assets and liabilities to construct a measure of the household's net wealth following Armantier, Armona, De Giorgi and van der Klaauw (2016).

[20]For a detailed overview of the SCE see Armantier, Topa, van der Klaauw and Zafar (2017).

Subjective Life Expectancy: The key reason for using the SCE here, rather than larger household finance surveys such as the Survey of Consumer Finances (SCF), is that households are also asked for their beliefs about their own longevity. Specifically, survey participants under the age of 65 are asked:

> "What do you think is the percent chance that you will live to age 85?"

This measures subjective survival beliefs. We do not take a stand on whether these beliefs are accurate or not. In our model we assumed that agents are correct about their survival probabilities, but the mechanisms are exactly the same if they are not: what matters for ambiguity-driven belief distortions and portfolio decisions is the agent's belief about their survival probability.

Alongside this question, survey participants are also asked for the probability that they will live to the age of 65 and 75. We focus on the longest-horizon question, as this contains the most cross-sectional variation, giving us the most power to detect our mechanism.[21] In Appendix C we repeat the regressions of Section 4.2 using these shorter-horizon questions, and find our results are qualitatively robust, though imprecisely estimated due to the smaller variation.

Summary Statistics: Table 1 displays summary statistics for the key variables in our analysis. In all cases, the statistics displayed concern the population who answered the subjective survival probability questions (i.e., those under 65).

Previous literature has highlighted in more detailed data that risky asset portfolio shares are typically increasing in age, income, and net worth. The binned scatter plots in Figure 6 show that this is also the case in the SCE, even though this data does not (for example) include indirect equity holdings through retirement accounts, which are typically included in portfolio data from the SCF.[22]

Discussion: All of the model results in Section 3 operate through distortions to beliefs about risky asset returns. We test our mechanism using data on portfolio shares, rather

[21] This is simply because almost all households in the sample expect to reach the age of 65 with a very high probability: the mean subjective survival probability to 65 in our sample is 82%. In contrast, the mean subjective survival probability to 85 is 53%. The variance in the answers to the 'age 85' question is more than double that of the 'age 65' question.

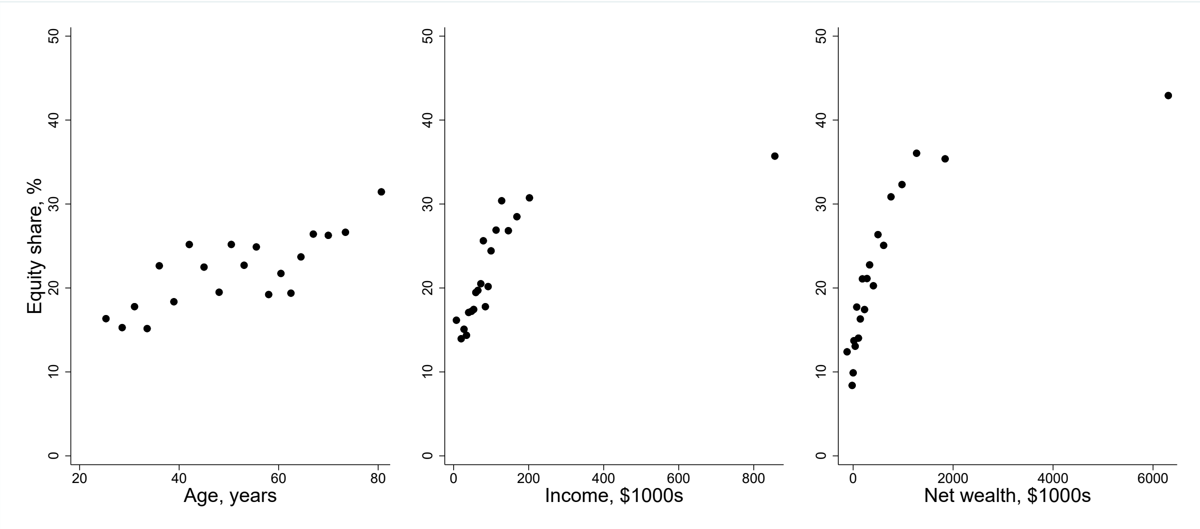
[22] Note that retirement accounts are still included in our measure of net wealth. They are only absent from our measure of the equity share, as specified in the survey question printed above. When studying the SCF, researchers typically construct risky asset shares by aggregating risky and safe components of portfolios, and thus can include equity holdings through retirement accounts (see e.g. Chang et al., 2018). That is not possible in the SCE data used here, as assets are not broken down to a sufficiently high level. The question on financial assets, in particular, asks households to combine equities, bonds, and various other assets. This is why we rely on the self-reported risky asset share, even though it excludes indirect holdings through retirement accounts.

Table 1: Summary statistics for key variables in the SCE Household Finance module.

	Mean	Median	Lower Quartile	Upper Quartile
Equity share (%)	20	0	0	35
Equity share, conditional (%)	45	40	20	70
Net wealth ($1000s)	503	139	6	460
Income ($1000s)	105	70	38	110
Subjective survival prob. (%)	52	50	30	75
Age (years)	45	46	35	56

Note: Summary statistics are computed for the sample of survey respondents with non-missing values for the subjective survival probability to age 85, which therefore only includes households with age < 65. "Equity share, conditional" refers to the equity share among the sample with non-zero shares of wealth invested in equity. Net wealth consists of the value of financial assets, housing, retirement accounts (IRAs and 401ks), other land, vehicles, and other assets, minus the value of mortgages, other home equity lines of credit, and non-housing personal debt (including credit card debt and student debt). Units for each variable are in brackets after the variable name. Sample period: 2014-2019. Source: Survey of Consumer Expectations, core survey and household finance module.

Figure 6: Share of financial assets invested in equities varies with age, income, and net wealth.



Note: Plots show binned scatter plots of the share of financial assets invested in stocks against age, income, and net wealth respectively. In each panel, each point represents the mean equity share and the mean x-axis realization in a given ventile of the x-axis distribution. Sample period: 2014-2019. Source: Survey of Consumer Expectations, core survey and household finance module.

than directly using return expectations, for two reasons.

First, in models of ambiguity aversion agents know that they are acting based on distorted beliefs. They simply act *as if* the worst case outcome is more likely than it really is. It is not immediately clear whether such agents would answer a survey about their return expectations with (i) the distorted beliefs they use for decision-making, or (ii) their initial undistorted beliefs, that still reflect their actual assessment of probabilities. Indeed, Adam, Matveev and Nagel (2021) find evidence that survey-based return expectations reflect undistorted beliefs. Since our mechanism only operates through distorted beliefs, this implies that survey-based expectations are not appropriate in this context.

Second, even if we dismiss the first conceptual problem, the asset return expectations elicited in the SCE are not well-suited to testing our particular mechanisms. The SCE elicits expectations of overall portfolio returns, but these combine expectations across a range of risky and safe assets, where our mechanism only concerns risky asset returns. The SCE also asks participants to give a probability that the US stock market will be higher in 12 months time than at the time of the survey, but this combines both a point expectation and the uncertainty in household beliefs, and so also cannot cleanly identify our mechanisms of interest (see Macaulay, 2021, for a discussion of this point).

4.2 Analysis

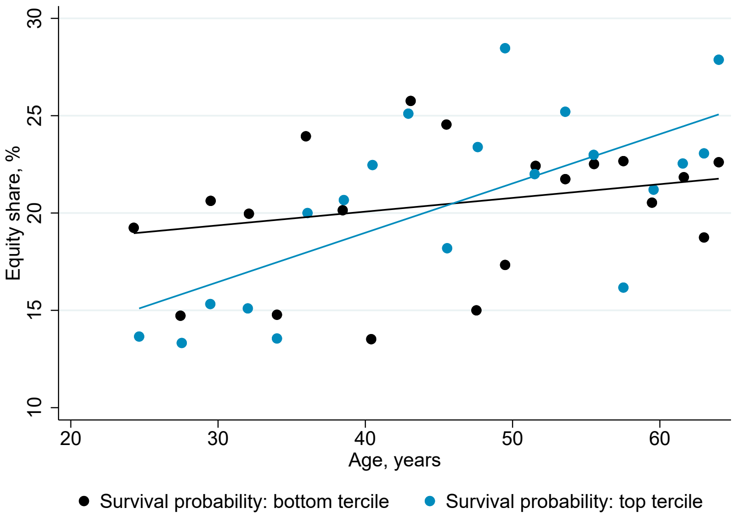
Corollary 2 predicts that when life expectancy increases, younger households become more pessimistic about risky asset returns, while older households become more optimistic – assuming a standard calibration in which $EIS < 1$. This is reflected in portfolio choices (Figure 3): increasing survival probabilities imply young households reduce the share of their wealth invested in risky assets, while middle-aged households increase their risky share. The age-gradient of risky asset shares therefore increases with survival beliefs.

This is the prediction that we test in the data. First, we simply plot the age-gradient of risky asset shares among those with high and low subjective survival probabilities. Figure 7 shows binned scatter plots of risky asset shares against age for those in the bottom tercile of the subjective survival probability distribution (black) and those in the top tercile (blue). Consistent with the model, a greater subjective survival probability is associated with lower risky asset shares among young households, but a steeper age gradient, such that greater longevity is associated with greater risky asset shares for older households.

Next, we test for this result in a more systematic way, and examine its statistical significance. For that we estimate the following equation via OLS:

$$\tilde{\alpha}_{it} = \beta_0 + \beta_1 age_{it} + \beta_2 \tilde{\phi}_{it} + \beta_3 age_{it} \cdot \tilde{\phi}_{it} + \delta_t + \Gamma' X_{it} + \varepsilon_{it} \tag{59}$$

Figure 7: Age-profile of portfolio share in equity, split by subjective survival probability.



Note: Plot shows the binned scatter plot of the share of financial assets invested in stocks against age, among those in the bottom ($\tilde{\phi}_{it} \leq 40\%$, black) and top ($\tilde{\phi}_{it} \geq 67\%$, blue) terciles of the subjective survival probability distribution. In each tercile, each point represents the mean equity share and the mean age in a given ventile of the age distribution. The solid lines give the predicted values from a linear regression of the equity share on age in each tercile, with equations $\tilde{\alpha}_{it} = 17.251 + 0.070 age_{it}$ and $\tilde{\alpha}_{it} = 8.857 + 0.253 age_{it}$ respectively. The slope coefficient is not significantly different from zero for the bottom tercile ($p = 0.317$) but is for the top tercile ($p = 0.00009$). Sample period: 2014-2019. Source: Survey of Consumer Expectations, core survey and household finance module.

where $\tilde{\alpha}_{it}$ is the share of household i's financial wealth invested in stocks in period t, age_{it} is their age in years, $\tilde{\phi}_{it}$ is their subjective probability of surviving to age 85,[23] δ_t are period fixed-effects, X_{it} is a vector of controls which we vary to check the robustness of the results, and ε_{it} is an error term.

The coefficients of interest are β_2 and β_3. Our model predicts that $\beta_2 < 0$ and $\beta_3 > 0$. The first of these inequalities implies young households invest less in equities when their life expectancy is longer, and the second implies that the age-gradient of equity shares is greater when life expectancy is longer.[24]

Table 2 shows the results. In column 1, the only controls are period fixed effects. Column 2 adds income and net wealth as extra controls, since the analysis in Corollary 2 and Figure 3 holds wealth (financial and human) constant. Column 3 adds a further range of demographic controls (details in table note). In all cases, the estimated coefficient on subjective survival probability (β_2) is negative, and the estimated interaction between subjective survival probability and age (β_3) is positive. Both are significantly different

[23]We use tildes here to distinguish these concepts from their model counterparts α_{it} and ϕ_{age}.

[24]More formally, the predicted change in $\tilde{\alpha}_{it}$ when $\tilde{\phi}_{it}$ rises for a given household is given by $\beta_2 + \beta_3 age_{it}$. $\beta_2 < 0$ and $\beta_3 > 0$ therefore correspond respectively to the results in Corollary 2 that young households reduce $\tilde{\alpha}_{it}$ and older households increase $\tilde{\alpha}_{it}$ after an increase in life expectancy.

from zero.

Table 2: Regressions on share of financial assets invested in equity.

	(1) Equity share	(2) Equity share	(3) Equity share
Age	-0.0919	-0.1189	-0.0745
	(0.0991)	(0.0989)	(0.0995)
Subjective survival prob.	-0.1659**	-0.1626**	-0.1467**
	(0.0745)	(0.0746)	(0.0718)
Age × Subjective survival prob.	0.0040**	0.0039**	0.0035**
	(0.0017)	(0.0017)	(0.0016)
Net wealth ($1000s)		0.0007*	0.0004
		(0.0004)	(0.0003)
Income ($1000s)		0.0023	0.0018
		(0.0017)	(0.0013)
Period FE	Yes	Yes	Yes
Demographic controls	No	No	Yes
Observations	3466	3446	3438
R^2	0.0054	0.0110	0.0724

Note: Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *All regressions are weighted using the SCE-provided survey weights. The demographic controls in column 3 consist of dummy variables for the state in which the respondent lives, gender, home ownership status, marriage status, race, and education. Sample period: 2014-2019. Source: Survey of Consumer Expectations, core survey and household finance module.*

The signs of the coefficients β_2 and β_3 are therefore consistent with the model. The magnitudes are also substantial. The 25th percentile of subjective survival probability is a household who believes they have a 30% chance of living to the age of 85. At this subjective life expectancy, using the estimates from the most demanding specification (Table 2 column 3) one extra year of age is associated with a 3 basis point increase in the unconditional risky asset share. At the 75th percentile of subjective survival probabilities (75% chance of surviving to 85), that gradient is more than 6 times steeper: an extra year of age is associated with a 19 basis point increase in the unconditional risky asset share.

Table 3 shows the results from the same analysis, estimated in the restricted sample for whom $\tilde{\alpha}_{it} > 0$. The dependent variable here is therefore the conditional risky asset share. Qualitatively, the results are the same as in Table 2.

These results are robust to a number of specification changes and checks, documented in the Appendix C. In particular, columns 2 and 3 of Tables 2 and 3 use income and

Table 3: Regressions on share of financial assets invested in equity, conditional on equity market participation.

	(1) Equity share	(2) Equity share	(3) Equity share
Age	-0.1935	-0.2115	-0.1112
	(0.1685)	(0.1688)	(0.1713)
Subjective survival prob.	-0.3444***	-0.3360**	-0.2625**
	(0.1323)	(0.1327)	(0.1323)
Age × Subjective survival prob.	0.0076***	0.0074***	0.0055*
	(0.0028)	(0.0029)	(0.0028)
Net wealth ($1000s)		0.0005*	0.0004*
		(0.0003)	(0.0002)
Income ($1000s)		0.0011	0.0011
		(0.0013)	(0.0010)
Period FE	Yes	Yes	Yes
Demographic controls	No	No	Yes
Observations	1579	1568	1562
R^2	0.0166	0.0181	0.0735

Note: Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *For other regression details see the note to Table 2.*

net wealth in levels, ensuring that we do not exclude households with zero or (for net wealth) negative values in these variables. However, restricting the sample to those with strictly positive income and net wealth and including these variables in logs does not meaningfully alter the coefficients of interest.

Importantly, this pattern is not predicted by any existing mechanism proposed in the literature. In particular, models in which beliefs depend only on age predict a constant *age profile* of return expectations (Campanale, 2011; Peijnenburg, 2018). In such models, it does not matter whether a 30-year old believes they will live a long life or not: their beliefs depend only on the fact that they are 30. In a Merton-style model such as the one we develop above, this would imply that risky asset shares conditional on age should be unaffected by life expectancy.

In fact, in richer models pure age-based mechanisms are even more at odds with the data. Households with longer life expectancy should invest more in equities, as their planning horizon is longer and so they can absorb more short-term risk (e.g. Yogo, 2016). Younger households should therefore have larger risky asset shares when their subjective life expectancy is high, which is the opposite of the patterns we observe in the data. Our model can therefore explain a fact in the data which would otherwise be extremely puzzling.

5 Quantitative Analysis

We now return to the model with J-period maximum lifespans from Section 2, and add the features now standard in quantitative life-cycle portfolio-choice models: risky age-dependent labor income, Epstein-Zin preferences, and equity market participation costs.[25] The resulting model is similar to that in Gomes and Michaelides (2005), with the addition of ambiguity aversion as specified in Section 2. As these quantitative model features are standard in this literature, we leave the formal details to Appendix D.1.

We calibrate the model demographics to survival probabilities in the US in 2019. We then use the model to examine the effect of demographic change on portfolios in the coming decades. To do this, we compare the 2019 calibration with a counterfactual using projected survival probabilities for 2100. Even holding the distribution of labor income fixed, changes in life expectancy imply non-trivial changes in the age profile of portfolio composition.

5.1 Calibration

We calibrate the model such that one period is one year. First, we set the majority of the parameters to typical values in the literature, as discussed in Gomes (2020) (see Appendix D.2 for details). We set the initial age to $j = 20$, and the maximum possible lifespan to $J = 109$: this is the first age in the 2019 mortality data discussed below at which the annual mortality rate exceeded 50%. The remaining parameters are the survival probabilities, and the two parameters not present in the model reviewed in Gomes (2020): the equity market participation cost, and the degree of ambiguity aversion (θ).

Survival probabilities: We first split households into three groups. For the first group, we calibrate their survival probabilities using age-specific mortality rates in the US in 2019 reported by the US Office of the Chief Actuary at the Social Security Administration. This mortality data is representative for the US population. The survival probability ϕ_j is one minus the empirical mortality rate for people of age j. The second group (high perceived mortality) have perceived mortality rates equal to $(1 - \phi_j)(1 + \zeta)$, while the third

[25] As in Cocco et al. (2005) and others, we do not include a bequest motive in our baseline exercises. However, in Appendix D.5, we extend the model to include a bequest motive, and find that our results are qualitatively unchanged.

group (low perceived mortality) have perceived mortality rates equal to $(1-\phi_j)(1-\zeta)$.[26] The perceived survival rates for each group are one minus the relevant perceived mortality rates.

This heterogeneity in perceived survival probabilities is a robust fact of our survey data (see Table 1). We therefore calibrate the scaling factor ζ to match that data. For each perceived-mortality group we calculate the implied perceived probability of surviving to age 85 at each age. We choose ζ so that the standard deviation of this perceived survival probability across agents below the age of 65 is equal to that in our SCE sample (28%).[27] All resulting mortality rates for all groups remain in $(0, 1)$.

The key reason for allowing heterogeneous perceived mortality rates is to provide a disciplined way to calibrate the degree of ambiguity aversion, as explained below. However, it is not crucial for our qualitative results. Appendix D.6 presents the key quantitative exercise of this section without any heterogeneity in perceived mortality rates, and our conclusions are robust. Throughout the rest of this section, we will compare the model with ambiguity to one without, but which is otherwise identical. That 'no ambiguity benchmark' replicates the standard results from other quantitative portfolio choice models, showing further than mortality rate heterogeneity is not driving any of our results.

Participation costs and ambiguity aversion: We calibrate the equity market participation cost to match the average participation rate in the 2019 Survey of Consumer Finances (53%). Finally, we calibrate the degree of ambiguity aversion to target the results of the empirical test of our mechanism in Section 4. Specifically, we target β_2 and β_3 in estimates of equation (59), as these coefficients form the key test of our ambiguity-driven mechanism.

Matching the regression coefficients: To target the results of Section 4, we estimate equation (59) on simulated data from the model.

Solving the model yields decision rules for all three mortality-rate groups. We simulate 10000 agents from each group, to give a simulation sample of 30000 agents, who differ by age, subjective probability of surviving to age 85, income, wealth, and portfolio choices. With this sample, we then estimate equation (59) on the simulated data for all agents below age 65 (to match the SCE sample). We use the conditional risky asset share as the dependent variable, because the unconditional share is more strongly affected by the

[26]This heterogeneity is in perceived mortality, not actual mortality. All agents' true mortality rates are the ones assigned to the first group, i.e., those taken from US data. All results except those in Appendix D.7 concern portfolio decisions conditional on age, so this distinction is largely not relevant, as portfolio decisions at each age depend on perceived and not actual mortality rates.

[27]We only consider agents below age 65 to match the SCE sample described in Section 4.1.

participation cost, which hinders our ability to separately identify θ from that cost.

This means that we replicate Table 3 in the model. We choose θ to target the key coefficients (β_2 and β_3) in the most demanding specification (column 3). With a plausible degree of ambiguity aversion (see Appendix D.3), we obtain $\beta_2 = -0.3073$, $\beta_3 = 0.0047$.[28] This slightly overstates the level effect of greater survival probability (β_2), and understates the interaction effect between age and subjective survival probability in particular (β_3), relative to the data. Since our main exercise below concerns the effect of increased longevity on the gradient of the age-profile of risky asset shares (Section 5.3), understating the interaction effect implies that our results in that projection are if anything somewhat conservative.

Note that this calibration approach is only possible because we have allowed for heterogeneous perceived mortality rates. Without that (i.e., if $\zeta = 0$), there would be no heterogeneity in subjective survival probabilities once we condition on an agent's age. In that case, we could not run the regression.

5.2 Age Profiles of Portfolio Allocations

Figure 8 plots the conditional risky asset share for agents of different ages in the calibrated baseline model, alongside the profile from an otherwise-identical model without ambiguity, and the empirical age profile from the SCE.[29] As in Section 4, we use $\tilde{\alpha}_{j,t}$ to refer to the share of risky assets in financial wealth, with the tilde to distinguish this from the share of risky assets in total wealth analyzed in Sections 2 and 3.

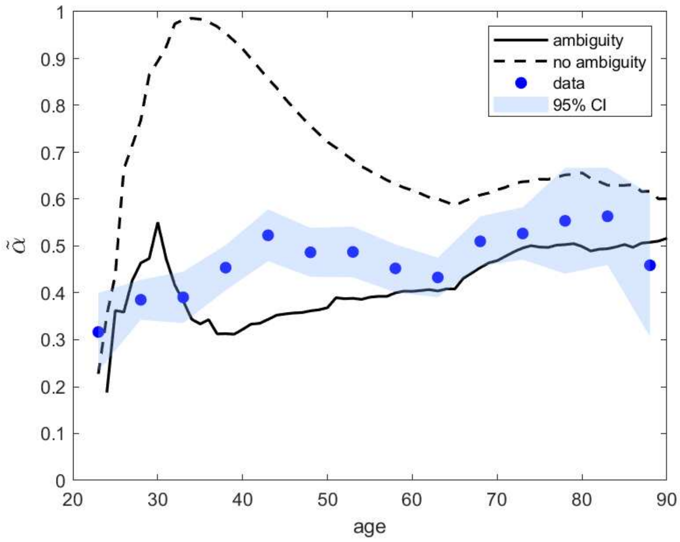
The calibrated model generates similar levels of conditional risky shares as seen in the data, and replicates the generally increasing age profile, despite neither of these being targeted in the calibration. Note that while the data displays a slight downward slope from ages 40-60, this is not systematically present in other datasets (see e.g. Chang et al., 2018; Catherine, 2022). The life-cycles of the participation rate and the unconditional risky asset share are in Appendix D.4.

Of course, the empirical age-profile shown is a snapshot of a particular point in time, and so conflates age, period, and cohort effects. A large literature attempts to disentangle the pure age effect using estimated models and a range of identifying assumptions (see Gomes and Smirnova, 2021, for a recent example). However, this is not our focus. A

[28]The match is not exact because we have one parameter and two targets. Increasing θ further increases β_2 (closer to target) but decreases β_3 (further from target). Our choice balances the two deviations in percentage terms.

[29]We use the SCE as this was used in the calibration step, but note that a similar age-profile has been observed in the Survey of Consumer Finances (Chang et al., 2018). The SCE data is pooled across 2014-2019 to obtain reasonable sample sizes in each age bin.

Figure 8: Model-generated vs. empirical conditional risky asset shares, 2019 calibration.



Note: The solid line is constructed using the model, with calibration described in Section 5.1 and Appendix D.2. The dashed line is constructed from the same model, with the preference for ambiguity set to $\theta = 0$. The circles and shaded areas are from the August waves of SCE from 2014-2019, as described in Section 4. Each circle is the mid-point of a five-year age range, and denotes the mean conditional risky asset share in that range. The shaded area is the 95% confidence interval around these means.

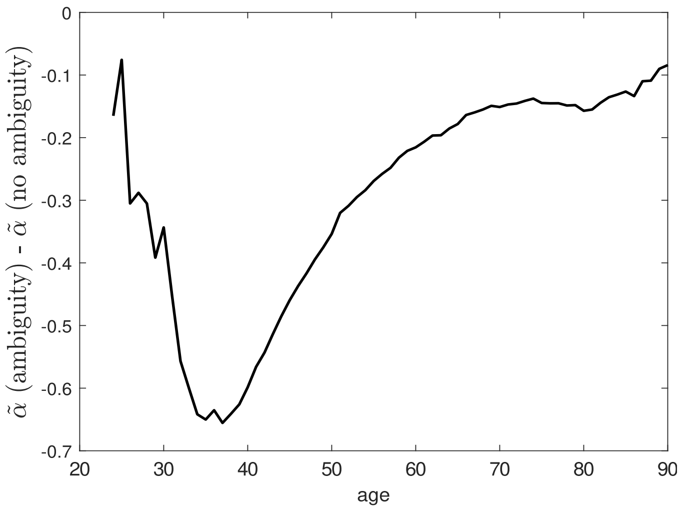
key part of our mechanism is that changes in life expectancy affect portfolio decisions, through changes in ambiguity-driven belief distortions. Since life expectancy changes over time and by cohort, we do not want to strip out these effects. Indeed, exploring how this profile might change over time is the purpose of Section 5.3.

To understand the mechanisms at work in this result, we begin by discussing the case without ambiguity, which is similar to many other standard models in this literature. In this benchmark, the conditional risky asset share rises rapidly early in life. This is a well-known effect of participation costs (see e.g. the discussions in Gomes and Michaelides, 2005; Catherine, 2022). After this point, the risky asset share decreases, then after retirement at age 65 it is mildly increasing. This path reflects the changing ratio of financial assets to 'human wealth' (the present value of future income). Since labor and retirement income are uncorrelated with risky asset returns, any expected future income acts like an extra endowment of risk-free bonds in the agent's portfolio problem. As the agent moves through their working life, their financial assets increase relative to the remaining expected future labor income. To maintain a constant share of risky assets out of total wealth, this shift to more financial wealth implies the agent must decrease the share of their financial assets invested in the risky asset. After retirement, wealth is gradually spent, implying that the financial wealth to income ratio decreases, reversing this trend. For a complete discussion of these well-known effects, see Gomes (2020) and the references therein.

The model with ambiguity features a similar path of risky asset shares at very young ages. However, the initial rise in α is much smaller, and the subsequent fall lasts for less time. After age 35, the risky asset share steadily increases, gradually converging back to the no-ambiguity benchmark.

To express this another way, Figure 9 plots the gap between the average conditional risky asset share in the benchmark no-ambiguity model and the model with ambiguity.[30]

Figure 9: The effects of ambiguity on conditional risky asset shares.



Note: The solid line is the difference between the 'ambiguity' and 'no ambiguity' lines in Figure 8. See the note to Figure 8 for details.

At very young ages, ambiguity has little effect. This is because the vast majority of very young agents' total wealth is contained in the present value of future labor income, not in financial wealth. Even though they have a high marginal utility of income, they are not therefore particularly exposed to fluctuations in financial portfolio returns.

However, outside of these very early years, the optimal belief distortions are similar to those implied by the simple model. In Figure 2, for the majority of the parameter space we found that young agents distorted their beliefs more than old agents. In the language of Section 3.2.1, the marginal utility channel dominates the wealth channel. As agents age from 35 to 65, their consumption rises (see Appendix D.4), and so their marginal utility of income falls, implying a smaller exposure to ambiguity in risky asset returns. The optimal belief distortion shrinks over these periods, reducing the effect of ambiguity. After retirement at 65, consumption does begin to fall, but more slowly than wealth. In

[30]Figure 27 in Appendix D.4 presents the same information with an alternative comparison, between the calibrated model and an alternative with a fixed distortion $\nu_{j,t}$ set so the two cases have the same average conditional risky share.

this region, the wealth channel therefore dominates, leading to further declines in belief distortions.

5.3 Projecting Asset Demand

We showed in Section 3 that life expectancy is a key driver of the age profile of portfolio choices. As populations age, survival probabilities increase particularly for older cohorts. In our model, this leads to differential changes in the portfolio choices of different age groups, with implications for inequality within and between cohorts.

To explore this dependence on demographics, we take the calibrated model and replace the 2019 survival rates with demographic projections for the year 2100 in the US.[31] As well as ϕ_j, we also recalibrate the maximum lifespan J using the same approach as before: we find the first age at which the annual mortality rate is expected to exceed 50%, which for 2100 implies $J = 117$. These projections come from the US Office of the Chief Actuary, who predict survival probabilities rising substantially, particularly for the oldest age groups.[32] This exercise therefore gives a projection of the direct effect of extending life expectancy on portfolio choices, holding everything else fixed.

Figure 10 shows the results, plotting the model-implied age profile of conditional risky asset shares in 2019 and 2100. Solid lines plot $\tilde{\alpha}_{j,2019}$ and dashed lines plot the equivalent $\tilde{\alpha}_{j,2100}$. The increase in life expectancy by 2100 causes younger middle-aged households to invest less in risky assets, as they distort their beliefs more strongly towards low risky returns. For older households, return expectations decline by less, so they become more optimistic relative to the young. To make these effects clearer, Figure 10 plots the difference between 2100 and 2019 for each age, alongside the equivalent "longevity effect" in the model without ambiguity.

Beyond very young agents, who as highlighted above are not very susceptible to this form of ambiguity, the age profile of risky asset shares therefore becomes steeper, in line with the results with $J = 2$ (Corollary 2).[33] Quantitatively, the gap between the

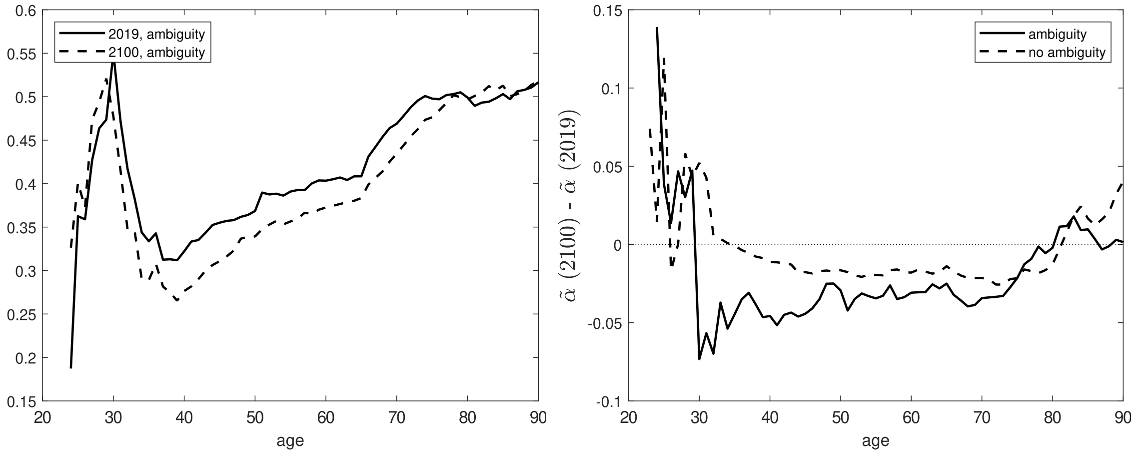
[31] A related exercise is performed in Auclert et al. (2021), which takes an OLG model and forecasts future wealth-to-income ratios by holding fixed the average wealth and income of each age group, but varying the proportions of households of each age according to UN projections. When we change ϕ_j, we keep the parameter governing heterogeneity in perceived mortality (ζ) constant.

[32] For example, in 2019 the death rate among 70-year-olds in the US was 1.9%. In 2100 that is projected to fall to 1.0%. These changes in death rates mean that life expectancy is projected to rise substantially. Life expectancy conditional on reaching age 30, for example, is projected to rise by more than 6 years, from 79.5 to 85.7. Conditional on reaching age 70, the projected rise is from 85.2 to 89.1. The data is available at https://www.ssa.gov/oact/HistEst/Death/2023/DeathProbabilities2023.html. Note that these projections contain substantial uncertainty, which we are abstracting from here. If life expectancy instead declines over this period, then the results would be the opposite of those we plot below, as the mechanisms outlined in Section 3 would operate in reverse.

[33] Note the participation rate is small for young agents in this model, so the simulated conditional

risky asset shares of those aged 80 and 35 rises from 16.6 p.p. to 20.9 p.p., an increase of 26%. There is no such effect in the model without ambiguity. If anything, older households decrease their risky asset shares slightly relative to younger households when facing increased longevity, which is counter to the cross-sectional evidence in Section 4.

Figure 10: Model-implied conditional risky asset shares, 2019 and 2100.



Note: Plots constructed using the calibration and data described in Section 5.1 and Appendix D.2. In panel (b), each line is the average conditional risky asset share at age j in the 2100 calibration, minus the equivalent average at that age in the 2019 calibration.

Plausible increases in survival probabilities therefore cause substantial changes in average portfolio decisions. Importantly, this projection only captures changes directly due to the age effect on ambiguity aversion and saving. If the income distribution or asset returns change over time, they would further alter these results. In Appendix D.7, for example, we explore one possibility, endogenizing the equity premium as in Section 3.5. In this case, the increase in life expectancy to 2100 causes the equity premium to rise, continuing its recent rising trend. However, we find that quantitatively the effect is modest, with the equity premium rising by 7 basis points, implying this particular channel does not substantially alter the conclusions presented here.

6 Conclusion

We develop a model in which investors face ambiguity over expected returns on risky assets. In contrast to previous literature, we allow agents to choose the degree to which they respond to this ambiguity optimally. With the same preferences, ambiguity aversion causes stronger distortions to return expectations among agents whose utility is very

risky share for those below the age of 30 is based on a small number of agents. This explains the noisy behavior of Figure 10 below age 30.

sensitive to risky asset returns. This implies differential effects of ambiguity on expected returns and portfolio choices for agents with different levels of wealth, and life expectancy.

In particular, as life expectancy rises, younger investors become more sensitive to rates of return, which means they distort their beliefs to be more pessimistic about the returns to risky assets, and they invest less in those assets. In contrast, in the empirically reasonable case where the elasticity of intertemporal substitution is less than 1, older agents become more optimistic about risky asset returns, and allocate a greater share of their savings to them. In this case wealthier households are also endogenously more optimistic about risky asset returns, fueling greater savings rates and risky asset shares, as documented in Straub (2019), Briggs et al. (2020) and others.

The prediction that greater life expectancy causes differential effects on the portfolio decisions of young and old households is novel to this model. We test it in survey data on US households, and find strong support: the age-profile of equity shares in portfolios is substantially steeper among households with a longer subjective life expectancy. This is difficult to explain without the ambiguity aversion we study.

In a quantitative extension of the model, we generate empirically plausible age profiles of risky asset shares in the US. We then use this quantitative model to project asset demand forward to 2100, to uncover the likely effect of demographic changes on savings behavior. As the population ages, and life expectancies increase, older households increase the share of their wealth invested in risky assets relative to the young. Given that the upward slope in the age profile of risky asset shares is already a puzzle in this literature, these results suggest that including ambiguity aversion in quantitative household finance models will become more and more critical in the coming decades.

To obtain these results, we have focused on a particular dimension of ambiguity that has received extensive empirical support (Dimmock et al., 2016). However, there are a range of related areas in which it is also plausible that agents face substantial ambiguity, which may themselves interact with life expectancy, and which could further alter financial decision-making. For example, agents may face ambiguity about their future income or survival probabilities (Caliendo, Gorry and Slavov, 2020), in addition to the ambiguity about their asset returns. Future research could profitably explore the interactions of these multiple dimensions of ambiguity, and how they are jointly affected by demographic change.

Chapter 2

Macroeconomic Impact of the Remote Work Revolution

Petr Sedláček
University of New South Wales

Chenchuan Shi
University of Oxford

Abstract

In recent years, remote work surged across the globe. We develop a novel macroeconomic model in which firms employ some workers remotely, trading-off potential productivity losses against savings on costs. Quantifying the model using U.S. data suggests that the surge in remote work (i) was primarily driven by workers' preferences, (ii) increased profitability and encouraged firm entry and (iii) shifted the firm distribution towards smaller businesses because they benefit relatively more from associated reductions in (fixed) costs. While increased entry sparks a boom, a larger share of small firms lowers aggregate productivity. Barriers to firm entry, therefore, emerge as a crucial margin determining the overall impact of the remote work revolution. Finally, we propose a novel identification strategy to estimate the firm-level effects of remote work, validating the model's key mechanisms and predictions.

JEL codes: D22, D24, E24, J23, L11, M13.

Sedláček (p.sedlacek@unsw.edu.au): University of New South Wales, Kensington, NSW 2052, Australia. Shi (chenchuan.shi@nuffield.ox.ac.uk): Department of Economics & Nuffield College, University of Oxford. We thank Christian Bayer, Nick Bloom, Timo Boppart, Andrea Chiavari, Steven Davis, Sergio De Ferra, Wouter den Haan, Martin Ellison, Andrea Ferrero, Michael Haliassos, Marek Ignaszak, Per Krusell, Alexandre Kohlhas, Kieran Larkin, Huiyu Li, Federico Masera, Gernot Mueller, Morten O. Ravn, Morgane Richard, Faisal Sohail, Emily Swift, Zhichun Wang, Francesco Zanetti and conference and seminar participants of the 2025 AEA meetings, T2M conference in Amsterdam, the Goethe University Frankfurt, the Computing in Economics and Finance conference in Singapore, the Nordic Summer Symposium in Macroeconomics, the Remote Work Conference, the University of New South Wales and the University of Oxford for useful comments. Sedláček gratefully acknowledges the financial support of the Australian Research Council (grant number FT230100545) and, in early stages of the project, of the European Commission and the European Research Council (grant number 802145).

1 Introduction

The COVID-19 pandemic sparked an unprecedented adoption of remote work arrangements. Fueled by forced experimentation, changes in attitudes towards remote work and new technologies, about one quarter of workdays occur remotely in the U.S. since the pandemic ended – more than five times the pre-pandemic average (see Barrero, Bloom and Davis, 2021, 2023). Similar values can also be found in other countries (see Aksoy, Barrero, Bloom, Davis, Dolls and Zarate, 2022). In this paper, we study the *macroeconomic* impact of this "remote work revolution."

Towards this end, we develop a novel macroeconomic model in which firms can choose to employ some of their workers remotely, optimally balancing the associated costs and benefits. On the one hand, remote work can reduce firms' production costs – e.g. because of a reduced need for production space (rent), or because workers prefer remote work and are willing to accept lower wages. On the other hand, remote work may come with decreased production efficiency (see e.g. Barrero et al., 2021, for a discussion).

To parameterize our model, we combine several U.S. (micro-)datasets on business dynamism, workers' time use and firm-level information on rental expenditures. In addition, we propose a novel identification strategy employing micro-data on firms' rental *commitments* prior to the pandemic with which we validate key model mechanisms and predictions. Using the parameterized model, we then study the main drivers and consequences of the remote work revolution. Two key messages stand out.

First, our model suggests that the dominant driver of the remote work revolution was a rise in preferences for working from home.[1] Second, our framework highlights that more favorable remote work conditions have two opposing macroeconomic effects. On the one hand, they raise overall firm profitability and, in turn, encourage business entry. This rationalizes the recently observed "surprising surge in applications for new businesses" (see Decker and Haltiwanger, 2024). On the other hand, however, the composition of firms shifts towards smaller businesses. This happens because smaller firms benefit relatively more from reductions in (fixed) costs brought about by cheaper remote work. While the former creates an economic boom, the latter lowers aggregate productivity.

The overall macroeconomic impact of the remote work revolution, therefore, crucially depends on the ease of firm entry. If a persistent rise in startups is hampered by financing or regulatory barriers, then weaker aggregate productivity can undo the individual (worker and firm) gains brought about by cheaper, more efficient and more preferred remote work. While recent U.S. data shows that firm entry has in fact remained persis-

[1] This result may also be interpreted as a post-pandemic change in social norms which placed remote work into the mainstream (see e.g. Aksoy et al., 2022).

tently elevated since the pandemic, evidence from other countries is mixed. Therefore, the welfare impact of the remote work revolution may differ substantially across the globe, depending on country-specific (barriers to) business dynamism.

More concretely, we begin our analysis by developing a core theoretical framework which allows us to *analytically* show how changes in remote work conditions affect business dynamism. In this model, individual firms – which differ in their (permanent) productivity levels – have the option of letting their employees work remotely. They do so by optimally balancing the associated costs and benefits.

On the one hand, remote work reduces costs. We consider two distinct reasons for this. First, workers may prefer remote work and, in return, accept lower wages (see e.g., Mas and Pallais, 2017; He, Neumark and Weng, 2021; Barrero et al., 2021). Second, remote work may lead to reductions in worker turnover and the associated training and hiring costs, or lower overhead costs because of a reduced need for production space (see e.g., Barrero, Bloom, Davis, Meyer and Mihaylov, 2022; Barrero et al., 2023; Bloom, Han and Liang, 2024). On the other hand, remote work may lower productivity. This can occur because of less efficient communication, mentoring and training or through reductions in worker motivation and self-control (see e.g., Natalia, Rune and Marieta, 2019; Battiston, Blanes i Vidal and Kirchmaier, 2021; Yang, Holtz, Jaffee, Suri, Sinha, Weston and Joyce et al., 2022; Emanuel and Harrington, 2023; Gibbs, Mengel and Siemroth, 2023).

Using our core model, we analytically show that more favorable remote work conditions have two opposing aggregate effects. Specifically, cheaper or more efficient remote work directly increases overall firm profitability. In the aggregate, this encourages firm entry. However, not all firms are affected equally. Small, less productive, businesses emerge as the "winners" of more favorable remote work conditions. For these firms, (fixed) costs represent a larger share of their expenditures and, therefore, they benefit relatively more when such costs are reduced through cheaper remote work.

To quantify which of these two effects eventually dominates, we generalize our core framework along several dimensions. First, we allow for fixed costs (heterogeneous across firms) of setting up remote work. This introduces an extensive margin, whereby only relatively productive (large) businesses can afford to start producing remotely. Note that this operates in the opposite direction to the intensive margin inherited from our core theory – *conditional* on producing remotely, smaller businesses tend to let more of their employees work from home. Second, we allow firm-level productivity to be affected by persistent idiosyncratic shocks and we *endogenize* the degree of long-run productivity differences across firms. Third, we introduce capital as a production factor and assume that its accumulation is subject to adjustment costs. Finally, we consider flexible labor supply, and explicitly model workers' preferences for remote work.

To parameterize the generalized model, we target (pre-pandemic) moments from several (micro-)datasets. First, we draw on the American Time Use Survey (ATUS) to compute remote work rates as the share of days worked from home among all work days. Second, we complement this information with the Current Population Survey (CPS) and its Annual Social and Economic Supplement (ASEC) in order to gain information on the size distribution of firms adopting remote work. Third, we use the Business Employment Dynamics (BED) data as a source of quarterly information on business entry, exit and size. Finally, we make use of firm-level information on rental expenses from Compustat to directly estimate one of the main cost margins intimately linked to remote work.

The parameterized model replicates salient features of the U.S. economy, with a particular focus on those pertaining to remote work. Specifically, our model matches average remote work rates overall and among large firms estimated from the ATUS and CPS. Moreover, the model does well in matching a range of *untargeted* empirical moments related to capital investment rates, firm-level productivity dynamics, the firm size distribution as well as the extent of productivity losses and cost savings associated with remote work.[2]

To quantitatively isolate the macroeconomic impact of the remote work revolution, we compare our baseline economy to a "remote economy" which is identical to the baseline but features more efficient, cheaper and more preferred remote work.[3] Recall from our core theory that all three of these forces incentivize firms to employ more workers remotely. However, savings on non-wage costs favor smaller firms relatively more. Therefore, to discipline the relative strength of (non-wage) cost reductions vs improvements in the efficiency of remote work, we require the remote economy to match the post-pandemic increase in work from home rates overall and over the firm size distribution. Finally, to discipline the importance of preferences for remote work, we makes use of existing estimates on the extent to which workers are willing to sacrifice wages in return for work flexibility (see e.g., Mas and Pallais, 2017; He et al., 2021).

As a first step in our quantitative analysis, we document that the dominant driver of the remote work revolution was a change in workers' preferences. In particular, our model suggests that stronger household preferences for working from home explain almost 2/3 of the surge in remote work rates following the pandemic. This result is consistent

[2]Our parameterization implies only small efficiency losses and cost savings for the average firm employing a fraction of its workers remotely. This is consistent with evidence that hybrid work arrangements may come with little to no efficiency losses (see e.g., Barrero et al., 2023).

[3]We do not consider transition dynamics between the baseline and remote economies since we view the pandemic period and the associated lockdowns as truly extraordinary. Instead, we compare the two stationary steady states because a sustained increase in remote work rates must ultimately be supported by underlying changes in the associated costs and benefits.

with recent evidence which also finds preferences to be key in understanding changes in remote work patterns (see e.g., Barrero et al., 2023; Bagga, Mann, Şahin and Violante, 2024; Zarate, Dolls, Davis, Bloom, Barrero and Aksoy, 2024).

Next, we use our model to study the aggregate impact of the remote work revolution. As highlighted by our core theory, more favorable remote work conditions come with increased profitability and, in turn, stronger incentives for firm entry. This is true also in the generalized model. Quantitatively, our model can explain almost 50 percent of the observed surge in firm entry. Importantly, however, increased firm entry raises labor demand and puts upward pressure on wages. This, in combination with reductions in the costs of remote work, creates winners and losers of the remote work revolution.

Specifically, the "winners" are smaller businesses conducting remote work which benefit relatively more from the associated (fixed) cost reductions. In contrast, the "losers" are larger firms operating only on-site. While these businesses do not directly benefit from more favorable remote work, they do feel the pain of a more competitive labor market. As a result, entry and exit decisions in the remote economy *endogenously* tilt the distribution of firms towards smaller businesses. Importantly, since small firms are on average less productive, this shift in the firm size distribution is also associated with lower aggregate productivity.[4]

Therefore, the boom driven by increased firm entry is slowed down by the endogenous shift of the economy towards smaller, less productive, firms. Which of these forces eventually prevails crucially depends on how strongly firm entry can increase. To highlight this, we consider a version of the remote economy in which barriers to entry (e.g., financial or regulatory frictions) mute the rise in startups. Our model suggests that, without a permanent surge in firm entry and associated economic boom, the welfare benefits of more favorable remote work can be entirely offset by a decline in aggregate productivity brought about by the endogenous shift towards smaller firms.[5] While recent data from the U.S. suggests that firm entry has in fact increased persistently since the pandemic, evidence from other countries is mixed. This suggests that the welfare impacts of the remote work revolution are likely to differ substantially across economies, depending on country-specific (barriers to) business dynamism.

As a final step in our analysis, we provide two pieces of empirical evidence in support of our key model predictions and mechanisms. First, we show that the model's predictions

[4]Not all small firms in our model are inefficient. Indeed, even productive firms start small and grow gradually over time. We discipline these firm-level dynamics by matching the life-cycle patterns of firm growth and exit observed in the data.

[5]While labor supply is flexible and workers have explicit preferences for remote work, we do not model additional benefits of remote work such as a decline in commuting time or benefits from home-production (see e.g., Barrero et al., 2023, for a discussion).

regarding firm entry, a shift in the size distribution of firms as well as changes in firm exit are all consistent with the data. This is true both qualitatively and quantitatively and both in the aggregate as well as across industries.

Second, we validate the key mechanisms of our model – i.e., we show that at the firm level, increases in remote work are associated with declines in firms' (rental) costs and labor productivity. A key empirical challenge is the lack of firm-level data on remote work. To overcome this, we propose a novel identification strategy utilizing Compustat data on firm-level *rental commitments*. In the data, some firms report having no rental commitments, while others report being committed several years into the future. To the extent that firms in 2019 did not predict the need for rental flexibility during the pandemic-induced surge in remote work, differences in rental commitments constitute exogenous variation in the *exposure* to the remote work revolution.

Intuitively, firms without rental commitments in 2019 were free to adjust their rental expenses during the pandemic if their workers (were forced to) work remotely. Moreover, because such firms could reap the cost-saving benefits of remote work, they were also more likely to *choose* higher remote work rates. Our model predicts that such firms should experience stronger post-pandemic declines in (rental) costs and labor productivity relative to firms with commitments. This is indeed borne out by the data. Moreover, a placebo treatment prior to the pandemic does not show any such effects. These results, therefore, provide additional independent validation of our model's key mechanisms.

Our paper is related to several strands of the literature. First, it contributes to research studying remote work (see e.g., Bloom, Liang, Roberts and Ying, 2015), with several very recent papers analyzing the (post-)pandemic period and focusing on household trade-offs, income and wealth, real estate prices, agglomeration economies and city structures (see e.g., Aksoy et al., 2022; Barrero et al., 2022, 2023; Davis, Ghent and Gregory, 2024; Decker and Haltiwanger, 2024; Hansen, Lamber, Bloom, Davis, Sadun and Taska, 2023; Liu and Su, 2024; Monte, Porcher and Rossi-Hansberg, 2023; Richard, 2024; Li, Sauvagnat and Schmitz, 2026). In contrast, we study the implications of remote work for business dynamism and, in turn, its *macroeconomic* impact. Second, we connect to the literature on the macroeconomic importance of heterogeneous firms – especially the influence of entry and exit (see e.g., Hopenhayn and Rogerson, 1993; Clementi and Palazzo, 2016; Sedláček and Sterk, 2017; Sedláček, 2020). Finally, we also link to a broader set of studies investigating the role of entry barriers in determining aggregate outcomes (see e.g., Poschke, 2010; Boedo and Mukoyama, 2012; Peters, 2020). To the best of our knowledge, we are the first to study remote work in these settings.

The rest of the paper is structured as follows. The next section lays out our core model and presents key theoretical results. Sections 3, 4 and 5 describe the generalized model, parameterize it and lay out our main quantitative results. Section 6 provides empirical evidence in support of our key findings and model mechanisms and the final section concludes.

2 Core Theoretical Framework

The main purpose of this paper is to study the influence of work from home patterns on business dynamism and, in turn, on the macroeconomy. In this section, we develop a tractable theory allowing us to derive analytical predictions and to build intuition. The next section generalizes our framework along several dimensions and brings it to the data in order to quantitatively evaluate the impact of remote work on the macroeconomy.

2.1 Model

Consider a framework with heterogeneous firms, indexed by j, each producing a final good sold to the household for consumption. A key novelty that we introduce in this section is the possibility of firms optimally choosing to employ workers remotely.

To ease the notation, we omit the (discrete) time index where possible and use upper-case letters to denote aggregates and lower-case letters for firm-level variables. For the purpose of this section, we focus only on firms' optimal decisions and we defer the remainder of the model, including a formal definition of its equilibrium, to the Appendix.

Production. To produce output, businesses use a common production function and combine labor, n_j, with firm-specific productivity, $z_j > 0$:

$$y_j = z_j n_j^{\alpha}, \tag{60}$$

where $\alpha \in (0, 1)$ denotes returns to scale and where firm-level productivity is assumed to be constant throughout firms' life-cycles.

Costs. In order to produce, firms must pay wages to their employees and a per-period operational fixed cost, κ_o. In addition, businesses also pay non-wage labor costs, κ_n, for every worker. Put together, total non-wage costs are given by $\kappa_n n_j + \kappa_o$.

The primary interpretation of these costs is as a continuous approximation to (likely staggered) expenditures on office or production space.[6] More broadly, these costs may also

[6]If adjusting office space incurred a cost, firms would optimally decide to increase office space in a staggered manner at certain firm size thresholds. Our cost structure, $\kappa_n n_j + \kappa_o$, serves as a continuous approximation to such an underlying step function.

represent expenditures on office or production equipment and supplies, worker training or hiring costs.

Work from home. We assume that all firms have the option of letting a fraction, $\omega_j \in [0, 1]$, of their employees work from home. This is associated with both costs and benefits, which we detail below and which the firm optimally balances.

Note that we abstract from the fixed costs of *setting up* remote work. We do so for tractability, but relax this assumption in the generalized model of Section 3. Therefore, the theoretical results in this section can be viewed as pertaining to the intensive margin of remote work, conditional on firms having paid a fixed setup cost.

Work from home: Wages. We assume that remote work directly helps alleviate wage pressures. This occurs because workers value time and location flexibility (see e.g., Mas and Pallais, 2017; Barrero et al., 2021, for experimental and survey evidence).

Therefore, we assume that a firm's wage bill falls as remote work rates increase. In particular, a firm's wage bill is given by $h(\omega_j)Wn_j$, where $h(\omega_j) \in (0, 1]$ with $h'(\omega_j) < 0$. W can then be viewed as the wage rate in firms that operate fully on site.

Work from home: Cost reductions. Aside from its impact on a firms' wage bill, remote work can also help reduce non-wage costs. This may occur because firms require less production or office space – see e.g., Barrero et al. (2023) for a discussion and Krause, Trumpp, Dichtl, Kiese and Rutsch (2024) for estimates of the post-pandemic decline in office space demand using German data.

In addition to production and office space expenditures, however, non-wage costs may also decline because remote work can reduce quit rates and the associated turnover and training costs (see e.g., Barrero et al., 2022). We model these effects by allowing (non-wage) labor and overhead costs to fall as remote work rates rise, $g(\omega_j)(\kappa_n n_j + \kappa_o)$, where $g(\omega_j) \in (0, 1]$, with $g'(\omega_j) < 0$.

Work from home: Productivity. While the previous two effects of remote work constitute benefits to the firm, work from home may also come with costs. In particular, we assume that producing with a larger fraction of remote workers can lower productivity.

Several studies show, in various settings, that fully remote work yields lower productivity than on-site work. These productivity losses of remote work occur because of impeded communication, less effective mentoring or management and reductions in worker motivation and self control (see e.g., Natalia et al., 2019; Battiston et al., 2021; Yang et al., 2022; Emanuel and Harrington, 2023; Gibbs et al., 2023; Liu and Su, 2024).[7] Therefore, we assume that firm productivity declines as remote work rates increase, $f(\omega_j)y_j$, where $f(\omega_j) \in (0, 1]$ with $f'(\omega_j) < 0$.

Combining the three effects of remote work on firms' operation – (i) wages, (ii) non-wage costs, and (iii) productivity – we can write firm profits as:[8]

$$\pi(z_j) = f(\omega_j)y_j - g(\omega_j)(\kappa_n n_j + \kappa_o) - h(\omega_j)W n_j. \tag{61}$$

Firm exit. All businesses are subject to an exogenous risk of shutting down, $\delta \in [0, 1)$.[9] Therefore, firm value is given by:

$$v(z_j) = \max_{n_j,\omega_j} \sum_{t=0}^{\infty} [\beta(1-\delta)]^t \pi(z_j) = \max_{n_j,\omega_j} \frac{\pi(z_j)}{1-\beta(1-\delta)}, \tag{62}$$

where $\beta \in (0, 1)$ is a discount factor.

Firm entry. There is a continuum of potential entrants which are, ex-ante, identical. In order to enter the economy, potential startups must pay a fixed entry cost, κ_e, upon which they obtain a draw of their (fixed) idiosyncratic productivity. Firms draw their productivity from a common distribution described by a probability and cumulative distribution function $h_z(z)$ and $H_z(z)$, respectively. Assuming free entry gives rise to the following entry condition:

$$\kappa_e = v_e, \tag{63}$$

where $v_e = \int v(z)h(z)dz$ is the expected value of entry.

[7]Studies of hybrid arrangements, i.e., partial work from home setups, find either no productivity effects or slight gains (see e.g., Bloom et al., 2015; Choudhury, Foroughi and Larson, 2021; Angelici and Profeta, 2023). While in reality firm-level productivity may rise for lower levels of ω before declining, in what follows we assume a monotone negative impact of remote work on productivity. This omission does not affect our results because – as will become clear – firms would always optimally choose levels of ω which imply productivity losses that exactly balance associated cost savings.

[8]Note that the effects of remote work on firms' operations can be interpreted as those *perceived* by firms. Changes in $f(\omega)$, $g(\omega)$ and $h(\omega)$ can then be viewed as learning about the uncertain impact of remote work. Indeed, studies and anecdotal evidence suggest that the COVID-19 pandemic lead to forced experimentation allowing a quick reduction in the previously high levels of uncertainty about the impact of remote work (see e.g., Barrero et al., 2021, 2023).

[9]Implicitly, we assume that fixed costs of operation, κ_o, and the distribution of firm-level productivity, z_j, are such that firms never choose to shut down endogenously. We relax this assumption in our generalized model.

2.2 Theoretical Results

In what follows, we study analytically optimal work from home choices, ω^*. In doing so, we pay special attention to how remote work choices differ with firm size and how they impact firm entry. We defer all proofs to the Appendix.

Optimal work from home. The following proposition summarizes firms' optimal work from home decisions and their relation to firm productivity. For simplicity, we assume $g(\omega) = h(\omega)$, allowing these to differ in our generalized model.

PROPOSITION **5** (Optimal work from home rates)
In the framework described above, for interior solutions and when $g(\omega) = h(\omega)$, optimal work from home rates, ω^*, satisfy the following

1. *if $\kappa_o = 0$, then ω^* is common across firms and implicitly given by*

$$\frac{f'(\omega^*)}{f(\omega^*)}\frac{g(\omega^*)}{g'(\omega^*)} = \alpha,$$

2. *if $\kappa_o > 0$, then*

$$\frac{\partial \omega^*}{\partial z} < 0.$$

The first part of Proposition 5 states that without fixed overhead costs, all businesses optimally choose the same level of work from home rates. Intuitively, firms choose work from home rates to balance the associated marginal cost (productivity declines) and benefits (cost savings). When remote work reduces wage and non-wage costs in the same way (i.e., when $g(\omega) = h(\omega)$), this trade-off mimics optimal labor demand. Therefore, in the absence of fixed overhead costs, optimal remote work rates are constant and common across firms, governed by returns to scale in production, α.

The second part of Proposition 5 constitutes one of our key results which will be crucial for the aggregate effects discussed in the later part of the paper. In particular, it states that with positive overhead costs, optimal work from home rates decrease with firm productivity. Intuitively, for less productive (smaller) firms, fixed overhead costs represent a larger share of their overall costs. This provides small firms with greater incentives to save on such costs by shifting more of their workforce off-site.

Note that, as explained above, our results in this section pertain to the intensive margin of remote work. In our generalized model, we allow for an extensive margin by introducing fixed setup costs of work from home. As will become clear, the extensive margin will work in the opposite direction to the intensive one, since larger businesses will be more readily able to pay the fixed setup costs. We address the quantitative question of which of these two forces dominates in Section 5.

Changes in work from home conditions. We now analyze how changes in work from home conditions affect firm profits and, in turn, entry decisions. Towards this end, let us denote $\widetilde{f}$ and $\widetilde{g}$ as parameters of $f(\omega)$ and $g(\omega)$ which, respectively, affect the speed of productivity losses and cost savings accrued with remote work. For simplicity, we continue to assume that $h(\omega) = g(\omega)$, i.e., $\widetilde{h} = \widetilde{g}$.

Without loss of generality, we define these parameters such that their increase leads to a rise in work from home rates:

$$\frac{\partial f(\omega;\widetilde{f})}{\partial \widetilde{f}} > 0, \quad \frac{\partial^2 f(\omega;\widetilde{f})}{\partial \omega \partial \widetilde{f}} > 0 \quad \text{and} \quad \frac{\partial g(\omega;\widetilde{g})}{\partial \widetilde{g}} < 0, \quad \frac{\partial^2 g(\omega;\widetilde{g})}{\partial \omega \partial \widetilde{g}} < 0.$$

Proposition **6** (Changes in remote work conditions)
All else equal and assuming internal optimal work from home rates, ω^*, exogenous changes in $\widetilde{f}$ and $\widetilde{g}$ have the following impact on firm entry incentives:

$$\frac{\partial v_e}{\partial \widetilde{f}} > 0 \ \textit{and} \ \frac{\partial v_e}{\partial \widetilde{g}} > 0.$$

Proposition 6 constitutes our second key result. Specifically, it states that entry incentives (summarized by the expected value of starting up a business, v_e) strengthen when remote work becomes cheaper or more efficient. Intuitively, such productivity boosts or cost reductions lead to higher profits, since businesses can produce more or at lower costs. In turn, higher profits, and hence firm values (62), incentivize firms to enter (63).

Before moving on, let us summarize the key mechanisms present in our core theory. First, conditional on conducting remote work, less productive (smaller) businesses benefit relatively more from the possibility of letting a fraction of their employees work from home – Proposition 5. Second, improvements in remote work conditions strengthen firm entry incentives – Proposition 6. As will become clear in the next section, these two effects are opposing forces which play a key role in shaping the overall macroeconomic impact of the remote work revolution.

3 Generalized Model

In this section, we generalize our core model along several dimensions. The next two sections then parameterize this model to U.S. data and use it as a laboratory to *quantitatively* evaluate which drivers were most important for the remote work revolution and how they impacted the macroeconomy.

The generalized model retains the structure of our core framework, but extends it along five important dimensions. First, we endogenize heterogeneous wages through employees' flexible labor supply and preferences for remote work. Second, we introduce the fixed costs (heterogeneous across firms) of setting up remote work. Third, we allow for endogenous firm exit. Fourth, we generalize firm-level productivity by (i) allowing it to be affected by persistent shocks and (ii) endogenizing the degree of long-run productivity differences across firms. Finally, in addition to labor, we also introduce physical capital as a production factor, the accumulation of which is subject to adjustment costs.

All these extensions have important consequences for the model-implied distribution of firms which, in turn, is what drives the responsiveness of the economy to structural changes – including the remote work revolution. Indeed, a primary use of our model will be to study the aggregate consequences of the remote work revolution. At this stage, it is important to stress that we will not use our framework to study aggregate fluctuations. Instead, our approach will rest on comparing steady-state equilibria which differ in the prevalence of remote work.

3.1 Household

We assume a representative household which owns all businesses in the economy and optimally chooses aggregate consumption and labor in individual firms.

Preferences for remote work. A key feature of household's utility is the preference for remote work, consistent with experimental and survey evidence (see e.g., Mas and Pallais, 2017; He et al., 2021; Barrero et al., 2021). Formally, we model the per-period household utility as:

$$\ln C - \sum_j \upsilon_j n_j, \tag{64}$$

where $\upsilon_j = \upsilon h(\omega_j) > 0$, with $h'(\omega) < 0$ as in our core model. In other words, while the household takes ω_j as given, its disutility of labor diminishes with if household members are allowed to work remotely.[10]

The representative household maximizes the expected present value of life-time utility subject to its budget constraint:

$$C = \sum_j W_j n_j + \Pi, \tag{65}$$

[10] We implicitly assume that all employees in a given firm work the same fraction, ω_j, of hours remotely.

where, normalizing the aggregate price level $P = 1$, real aggregate profits are given by Π and firm-level real wages are given by W_j. Wages are heterogeneous across firms because businesses optimally choose different levels of remote work – partly driven by the fact that in return for flexibility workers are willing to accept lower wages according to the following labor supply condition:

$$W_j = \upsilon_j C. \tag{66}$$

3.2 Firms

In the absence of aggregate uncertainty, all aggregates will be fixed at their respective steady state values. However, firm-level variables will in general fluctuate over time, reflecting changes in firm-specific (endogenous and exogenous) state variables. Therefore, whenever necessary, we denote time with a subscript t.

Costs of setting up remote work. As in our stylized model, work from home is associated with efficiency losses in production, summarized by $f(\omega_{j,t})$, and reductions in non-wage costs, summarized by $g(\omega_{j,t})$.

An important novel feature of our generalized model is the presence of firm-level costs of *setting up* remote work. These setup costs may represent not only costs of hardware and software necessary for remote work, but also the costs associated with developing and implementing efficient protocols and procedures for remote communication. We denote these fixed costs as κ_j^ω and allow them to be heterogeneous across firms (but fixed over time).

Every period, firms decide whether or not to pay the fixed setup costs. If a business decides not to pay the setup cost, it cannot employ workers off-site and, therefore, $\omega_{j,t} = 0$. Once a business pays κ_j^ω, it has the option to employ workers remotely in all future periods, i.e., $\omega_{j,t} \in [0, 1]$. We discuss all optimal firm decisions below.

Productivity process. We assume that firm-specific productivity, $z_{j,t}$, evolves according to the following law of motion:

$$\ln z_{j,t} = \underline{z}_j(1 - \rho) + \rho \ln z_{j,t-1} + \epsilon_{j,t}, \tag{67}$$

where $\rho \in (0, 1)$ is the persistence of firm-level productivity and ϵ_t are productivity shocks which are distributed identically and independently across firms and over time according to the distribution function H_z with zero mean and dispersion σ_z.

The parameter $\underline{z}_j$ represents the unconditional, long-run, mean of firm-level productivity. We assume that $\underline{z}_j$ is heterogeneous across firms but fixed over time.

Permanent heterogeneity. At this point, we highlight that firms in our model are characterized by *permanent* differences. These are governed by differences in firms' long-run productivity, $\underline{z}_j$, and remote work setup costs, κ_j^ω. In what follows, we will refer to these permanent differences as firm "types" and we will describe in detail how type heterogeneity is determined in our model.

It is important to note that different types of firms will make different decisions, even conditional on the same firm-level state variables. To ease the notation, we will not explicitly denote firm-level choices with $\underline{z}_j$ and κ_j^ω unless necessary for the clarity of our exposition. Instead, we will use the firm-level subscript j, implicitly understanding that each firm is characterized by its own pair of permanent characteristics $\underline{z}_j$ and κ_j^ω.

Production. Firms produce output using labor, $n_{j,t}$, and capital, $k_{j,t}$. They do so according to the following production function:

$$y_{j,t} = f(\omega_{j,t}) z_{j,t} \left(n_{j,t}^{\alpha} k_{j,t}^{1-\alpha} \right)^{\theta}, \tag{68}$$

where $\alpha \in (0,1)$ and $\theta \in (0,1)$ are common across all firms. As mentioned above, the efficiency of production is affected by firms' work from home decisions according to $f(\omega_{j,t})$.

Capital adjustment costs. We assume that firms accumulate capital subject to adjustment costs. In particular, investing $x_{j,t}$ into capital accumulation comes at a cost $\zeta(x_{j,t}, k_{j,t})$. The stock of firm-level capital then evolves according to the following law of motion:

$$k_{j,t+1} = x_{j,t} + (1 - \delta_k) k_{j,t}, \tag{69}$$

where $\delta_k \in (0,1)$ is the capital depreciation rate and where we assume that capital becomes productive only in the next period.

Fixed operational costs. As in the core theoretical model, firms must pay a per-period fixed overhead cost, κ_o. However, in our generalized framework, we assume that these costs are stochastic, distributed identically and independently over time and across firms according to the cumulative distribution function H_κ with mean μ_κ and dispersion σ_κ. As will become clear, it will be convenient to denote the stochastic component of overhead costs as $\widetilde{\kappa}_o = \kappa_o - \mu_\kappa$, where $\widetilde{\kappa}_o$ is distributed according to H_κ with zero mean and dispersion σ_κ.

In contrast to the core model, our generalized framework allows for the possibility of endogenous firm exit. This happens whenever the firm value falls below zero. We next turn to describing all optimal firm decisions.

Optimal decisions of incumbent firms. Every period, incumbent firms choose whether or not to stay in operation and – if they decide to continue – how many workers to hire and what amount of resources to devote to capital accumulation. In addition, businesses in our framework must also choose what fraction of their employees to conduct remote work. Before being able to do so, however, they must first pay the fixed cost of setting up remote work.

Formally, businesses make their decisions to maximize the net present value of current and all future profits. In particular, the beginning-of-period value of a business in operation which has *not* yet paid the fixed setup cost of setting up remote work is given by

$$v_j(z_{j,t}, k_{j,t}) = \max_{n_{j,t}, x_{j,t}} \left\{ \pi_{j,t} + \beta(1-\delta) u_j(z_{j,t}, k_{j,t}) \right\}, \tag{70}$$

where $\pi_{j,t} = f(\omega_{j,t}) y_{j,t} - W(\omega_j) n_{j,t} - g(\omega_{j,t})(\kappa_n n_{j,t} + \mu_\kappa) - x_{j,t} - \zeta(x_{j,t}, k_{j,t})$ are per-period profits and where we have made explicit that firm-specific wages depend on remote work. In particular, using the household's labor supply decision (66), we can write $W(\omega_j) = W_j = \upsilon h(\omega_j) C$. Recall that firms which have not yet paid the fixed setup cost of remote work cannot choose to have part of their employees work from home. Therefore, for these firms $\omega_{j,t} = 0$.

In the above, u_j is the continuation value of a business which is not yet doing remote work. This continuation value summarizes the optimal choice between three options: (i) shutting down, (ii) continuing purely on-site or (iii) paying the fixed setup cost and continuing as a firm which can produce remotely. Formally, the continuation value is given by

$$u_j(z_{j,t}, k_{j,t}) = \int \max \begin{bmatrix} v_j^x(k_{j,t+1}), \mathbb{E}_t v_j(z_{j,t+1}, k_{j,t+1}) - \widetilde{\kappa}_o, \\ \mathbb{E}_t v_j^\omega(z_{j,t+1}, k_{j,t+1}) - \widetilde{\kappa}_o - \kappa_j^\omega \end{bmatrix} dH_\kappa(\widetilde{\kappa}_o), \tag{71}$$

where $\mathbb{E}$ is an expectation operator with respect to the evolution of firm-level productivity. The exit value, v^x, is given by

$$v_j^x(z_{j,t}, k_{j,t}) = k_{j,t}(1-\delta_k) - \zeta(-k_{j,t}(1-\delta_k), k_{j,t}), \tag{72}$$

where firms obtain value from selling their stock of capital, but have to take into account the adjustment costs of doing so. The value of a firm which has paid the setup cost for remote work is given by

$$v_j^\omega(z_{j,t}, k_{j,t}) = \max_{n_{j,t}, \omega_{j,t}, x_{j,t}} \left\{ \pi_{j,t} + \beta(1-\delta) u_j^\omega(z_{j,t}, k_{j,t}) \right\}, \tag{73}$$

where the continuation value now contains only two options: (i) exit or (ii) stay in business with the possibility of doing remote work. Formally, the continuation value is given by:

$$u_j^\omega(z_{j,t}, k_{j,t}) = \int \max\left[v_j^x(k_{j,t+1}), \mathbb{E}_t v_j^\omega(z_{j,t+1}, k_{j,t+1}) - \widetilde{\kappa}_o\right] dH_\kappa(\widetilde{\kappa}_o). \tag{74}$$

Note that firms must pay the fixed setup cost of remote work only once. After it has been paid, firms do not have to pay it again even if they decide not to conduct remote work at times but "restart" again in later periods, that is, when $\omega_{j,t} = 0$ but $\omega_{j,t+s} > 0$ for $s > 0$.

Entry. Recall that there are permanent differences across firms, summarized by the subscript j. For the purpose of describing firm entry, however, we will make explicit the dependence of firms' decisions on the underlying parameters $\underline{z}_j$ and κ_j^ω. In particular, let $v^\omega(z, k; \underline{z}, \kappa^w)$ and $v(z, k; \underline{z}, \kappa^w)$ be the firm value of a business with productivity z, capital stock k, long-run productivity $\underline{z}$ and a remote work setup cost κ^ω which, respectively, has and has not paid the fixed setup cost.

For tractability, we assume a finite number of different productivity and fixed cost types. Specifically, let I be the number of different long-run productivity types and L the number of different setup cost types. The distribution of firm types is endogenous, modeled along the lines of Sedláček and Sterk (2017).

In particular, potential startups are free to choose which type of long-run productivity businesses they will *attempt* to start up. In order to do so, they must first pay an entry cost, κ^e, common across business types. This allows them to compete for a limited and time-invariant number of business opportunities of a given productivity type, denoted by Ψ_i.

Each business opportunity is exclusive, allowing for at most one producer. This means that not all potential startups succeed if multiple competitors attempt to seize a single opportunity. Specifically, the mass of successful startups of a given productivity type, m_i, is determined by the following "entry function"

$$m_i = \Psi_i^\phi s_i^{1-\phi}, \tag{75}$$

where s_i is the mass of startup attempts of type i and $\phi \in (0, 1)$ determines the degree of crowding out which is common across productivity types.

Upon entry, firms are randomly (and independently from productivity types) assigned a fixed setup cost of remote work, κ_l^ω. We use p_l to denote the probability of a particular cost type, where $\sum_l p_l = 1$. Therefore, assuming free entry, we obtain the following entry conditions

$$\kappa_e = \frac{m_i}{s_i} \sum_l p_l \int_z \max \left[v(z, 0; \underline{z}_i, \kappa_l^\omega), v^\omega(z, 0; \underline{z}_i, \kappa_l^\omega) - \kappa_l^\omega \right] dH_z(z), \tag{76}$$

where we assume that firms enter with zero capital and an initial productivity draw from the distribution $H_z(z)$. The total mass of entrants is then given by $M = \sum_i m_i$. Notice that in equilibrium, potential startups are indifferent between business types. This happens because business types with high expected payoffs (firm values) attract more startup attempts which, in turn, lowers the chances of successfully starting up.

Finally, note that since firm entry is determined by expected firm values, the mass of entrants of any given type is constant in the absence of aggregate uncertainty. However, while constant in the stationary steady state, the distribution of firm types is *endogenous*. Importantly, for purposes of this paper, our model allows for the possibility that changes in work from home conditions will influence this distribution of startup types.

3.3 Aggregation

Let $\mu_{i,l}(z, k)$ denote the distribution of firms with long-run productivity $\underline{z}_i$ and setup costs κ_l^ω across productivity levels, z, and capital holdings, k. Then, the following expressions describe goods and labor market clearing:

$$Y = \sum_i \sum_l \int \int y \; \mu_{i,l}(z, k) dz dk, \tag{77}$$

$$N = \sum_i \sum_l \int \int n \; \mu_{i,l}(z, k) dz dk. \tag{78}$$

Finally, the aggregate resource constraint is given by

$$Y = C + S\kappa_e + \sum_i \sum_l \int \int \left[\zeta(x, k) + g(\omega)(\kappa_n n + \mu_\kappa) + \widehat{\kappa}_o + \kappa_l^\omega \mathbb{1}_{i,l}(z, k) \right] \mu_{i,l}(z, k) dz dk, \tag{79}$$

where $S = \sum_i s_i$ is the total mass of startup attempts and where aggregate output is used for consumption and all paid costs. The latter include costs of firm entry and capital adjustment, non-wage labor costs, overhead costs (where $\widehat{\kappa}_o$ indicates the *paid* overhead costs, conditional on firm survival) and setup costs of remote work. For the latter, $\mathbb{1}_{i,l}(z, k)$ is an indicator function which is equal to one if a firm with long-run productivity $\underline{z}_i$, setup costs κ_l^ω and productivity and capital z and k, respectively, decides to pay the setup cost and zero otherwise. We defer a formal definition of the equilibrium to the Appendix.

4 Taking the Model to the Data

In this section we bring our generalized model to the data. In doing so, we ensure that it is consistent with a range of salient features of the U.S. economy pertaining to business dynamism and remote work arrangements. Our baseline calibration uses a model period of one year and targets pre-pandemic (2003-2019) moments of the U.S. economy. The next section describes in detail how we quantitatively isolate the macroeconomic impact of the remote work revolution.

4.1 Data

To parameterize our model, we use information from five different data: (i) the Business Response Survey, (ii) the American Time Use Survey, (iii) the Annual Social and Economic Supplement of the Current Population Survey, (iv) Business Employment Dynamics and (v) Compustat. In what follows, we briefly describe each dataset and which set of targeted moments it is used for.

Work from home. When analyzing remote work, we focus on both the share of hours worked remotely and the share of establishments conducting remote work. To measure the latter, we rely on the Business Response Survey (BRS) of the Bureau of Labor Statistics (BLS) which started in 2020. The survey offers, among other things, information on the fraction of establishments conducting remote work, including just prior to the pandemic.

To measure the intensive margin, we rely on the American Time Use Survey (ATUS), also conducted by the BLS. The ATUS provides monthly information (starting in January 2003) on how individuals in the U.S. allocate their time among various activities. The sample of households is connected to the Current Population Survey (CPS) allowing us to link individuals' time allocation data to other information, such as the industry they work in.[11] In addition, utilizing the Annual Social and Economic Supplement (ASEC) of the CPS allows us to infer the size of establishments for which individuals in the ATUS report working remotely.

Business dynamism. To measure the entry and exit of businesses, we use the Business Employment Dynamics (BED) dataset of the Bureau of Labor Statistics. This dataset is generated from the Quarterly Census of Employment and Wages (QCEW) and offers quarterly information on employment at the establishment level covering approximately

[11] The ATUS targets households which have completed their final (eighth) month of the CPS. From each of the selected households, a random individual aged 15 and over is chosen to participate in ATUS. The questionnaire asks information about the respondent's previous day and is conducted only once for each individual. For more details on ATUS, see Hamermesh, Frazis and Stewart (2005).

98 percent of all employment in the U.S. economy.[12] A key advantage of this dataset is its relatively timely nature with the latest data – at the time of writing this paper – running all the way to Q4 2022 allowing us to analyze the post-pandemic period.

Establishment entry – formally called "births" in the BED – is defined as units which record positive employment for the first time in a given quarter and which exclude (seasonal) re-openings of businesses. Symmetrically, establishment exit – formally called "deaths" in the BED – is defined as units with zero employment which exclude temporary closings of businesses.[13]

Finally, the BED does not report overall establishment size at a quarterly frequency. However, for establishment births and deaths, it can be imputed using information on the number of entering or exiting establishments and their respective employment levels. Therefore, in our analysis we focus on the size of entrants and exiters instead of average size of all establishments.[14]

Non-wage labor costs. As already highlighted in our theory, firm-level costs play an important role for understanding the heterogeneous effects of changes in remote work conditions. Firms' rental expenses are one of the major costs directly affected by remote work choices. For this reason, we make use of Compustat which offers detailed information firms' rental expenses.

While Compustat is one of the main sources of U.S. firm-level information and has been used extensively in economic research it is well known that the Compustat sample of (publicly traded) firms is not representative. To address this issue, when computing model-generated moments with counterparts in Compustat, we use estimated size-based weights which align the Compustat and BED firm-size distributions (see the Appendix for details).

[12]The BED excludes self-employed individuals, government institutions and some non-profit organizations. An alternative popular data source for business dynamism is the Business Dynamic Statistics (BDS) of the Census Bureau. While there exist differences between the BDS and the BED, the numbers of establishments as well as their employment sizes typically co-move strongly across the two datasets (see e.g., Decker and Haltiwanger, 2024, for a discussion). In the Appendix, we provide a comparison between the BDS and the BED showing that for our purposes they are similar in the overlapping periods.

[13]To determine whether a shut down is a death or temporary closure, the BLS requires establishments to report zero employment for four consecutive quarters before it classifies it as a death. Such establishment deaths are then "back-dated" to the relevant quarter when they occurred. Moreover, the Appendix shows that our results are similar when using establishment openings and closings as opposed to the stricter births and deaths.

[14]The Appendix shows that our results are similar when using overall establishment size imputed from the QCEW – the underlying source for the BED which is available quarterly but is, however, based on a somewhat different sample (see `https://www.bls.gov/opub/hom/cew/concepts.htm`) – or *annual* establishment size taken from the BED.

4.2 Parameterization

In what follows, we describe our parameterization strategy. This entails explaining our functional form choices and how we set parameter values in order to match key targets from our datasets. All model parameters are summarized in Table 4.

Functional forms and permanent firm heterogeneity. To bring our model to the data, we need to assume particular functional forms for the remote work productivity loss, non-wage cost saving functions and preferences for remote work, $f(\omega)$, $g(\omega)$ and $h(\omega)$. Towards this end, we follow León-Ledesma and Satchi (2019) in their analysis of technology adjustment and specify f, g and h as versions of exponential functions:

$$f(\omega) = \exp\left(\widetilde{f}\omega^2\right), \quad g(\omega) = \exp\left(-\widetilde{g}\omega\right), \quad h(\omega) = \exp\left(-\widetilde{h}\omega\right). \tag{80}$$

Note that, without loss of generality, the above specifications imply that increases in $\widetilde{f}$, $\widetilde{g}$ and $\widetilde{h}$ all result in remote work becoming more favorable or desirable.

Next, we follow Cooper and Haltiwanger (2006) and assume the following capital adjustment costs

$$\zeta(x, k) = \zeta_0(x)k + \frac{\zeta_1}{2}\left(\frac{x}{k}\right)^2 k, \tag{81}$$

where $\zeta_0(x) = \zeta_0$ whenever investment, x, is non-zero and $\zeta_0(x) = 0$ otherwise.

Recall that our generalized model features permanent firm heterogeneity in terms of long-run productivity and setup costs of remote work. Therefore, aside from the above functional form choices, we also need to make a stand on the parameterization of permanent firm-level differences.

For tractability, we assume two types of firms along each dimension, indexed by subscripts L and H to denote "low" and "high" types. Long-run differences across firms are then governed by seven parameters: four level parameters ($\underline{z}_L$, $\underline{z}_H$, κ_L^ω and κ_H^ω), two parameters controlling the masses of low and high productivity type startup opportunities ($\Psi_L > 0$ and $\Psi_H > 0$) and the share of low setup cost firms which we denote by $\Psi_\omega \in [0, 1]$.

Common choices and normalizations. We set the discount factor to $\beta = 0.96$, reflecting a roughly 4% annual interest rate. The production function parameters are given by $\alpha = 0.65$ and $\theta = 0.9$. While the former mimics the observed labor share in income, the latter falls within the span of control values estimated in the data and commonly used in the literature (see e.g., Basu and Fernald, 1997; Clementi and Palazzo, 2016). We set the capital depreciation rate at 8% per year which lies in between values used in the literature (see e.g., Cooper and Haltiwanger, 2006; Clementi and Palazzo, 2016).

We set the disutility of labor υ such that the wage rate in on-site only firms is normalized to $W = 1$. Similarly, we assume the entry cost κ_e is such that the mass of entrants is normalized to $M = 1$. Following Sedláček and Sterk (2017), we set $\phi = 0.156$ and provide robustness exercises in the Appendix.

Indirect inference. The remainder of the parameters are set to match a range of business dynamism and work from home moments in the data. While all model parameters affect the behavior of the entire model, we discuss the targeted moments in relation to the parameters to which they are tied the most. Specifically, there are 17 remaining parameters: the persistence and dispersion of productivity shocks, the two long-run means and the respective masses of business opportunities (ρ, σ_z, $\underline{z}_L$, $\underline{z}_H$, Ψ_L, Ψ_H), the mean and dispersion of fixed overhead cost (μ_κ, σ_κ), capital adjustment cost parameters (ζ_0 and ζ_1), parameters controlling the impact of remote work on productivity, non-wage costs and preferences ($\widetilde{f}$, $\widetilde{g}$, $\widetilde{h}$), the level of non-wage labor costs (κ_n), the level of remote work setup costs (κ^ω_L, κ^ω_H) and the respective fraction of startups with low setup costs, Ψ_ω.

Practically, we compute the selected model-generated moments and compare them with their respective empirical counterparts and minimize the following loss function:

$$\min \sum_m \left(\frac{\text{model(m)} - \text{data(m)}}{\text{data(m)}} \right)^2,$$

where m indicates a given moment. Note that our model is over-identified as we are estimating 17 parameters using 26 moments described below. Details of the computational strategy are provided in the Appendix.

Indirect inference: Remote work setup costs. Recall that there are two types of firms when it comes to the costs of setting up remote work. The distribution of these firm types is governed by the share of low cost firm types (Ψ_ω) and the respective levels of setup costs (κ^ω_L and κ^ω_H).

First, we assume that the low remote work setup cost is $\kappa^\omega_L = 0$. This reflects the fact that for some businesses the necessary hardware and software for conducting remote work is part of their regular operations (i.e., subsumed in their capital stock) and that basic versions of remote work telecommunications services are often available free of charge.

Given κ^ω_L, we pin down κ^ω_H and Ψ_ω by targeting the pre-pandemic fraction of firms conducting remote overall and among large firms as reported in the Business Response Survey. Intuitively, since the minimum setup cost is $\kappa^w_L = 0$, all firms with such costs will choose to "pay" them. Therefore, the fraction of firms conducting remote work is informative about Ψ_ω. In contrast, only relatively productive (large) firms are capable

Table 4: Parameter values and targeted moments

	Parameter	Value	Target/Source	Data	Model
β	Discount factor	0.96	Interest rate of approx. 4%		
α	Labor elasticity of output	0.65	Labor share in income of approx. 65%		
θ	Returns to scale	0.90	Basu, Fernald (1997) estimate		
δ_k	Capital depreciation	0.08	Cooper, Haltiwanger (2006)		
υ	Disutility of labor	0.017	Normalization, $W = 1$		
κ_e	Entry cost	0.67	Normalization, $s_H + s_L = 1$		
ϕ	Elasticity of entry function	0.156	Sedláček and Sterk (2017)		
κ_L^ω	"Low" remote work setup costs	0	Normalization, minimum remote work setup cost of 0		
$\widetilde{f}$	Speed of remote work productivity losses	-0.151	Average work from home rate, ATUS	4.1%	3.9%
$\widetilde{g}$	Speed of remote work cost savings	0.325	Work from home rates in 100+ over 100- firms, ATUS & ASEC	1.12	1.12
$\widetilde{h}$	Household preferences for remote work	0	Wages unresponsive to remote work	0	0
κ_n	Non-wage labor costs	0.255	Rental expenses - size coefficient, Compustat (see (83))	0.86	0.91
κ_H^ω	"High" remote work setup costs	5.3	Share of large firms conducting remote work, BRS	30%	29%
Ψ_ω	Share of "low" remote work setup costs	0.195	Share of firms conducting remote work, BRS	23%	23%
Ψ_L	Mass of "low" business opportunities	$1.1e-4$	Share of small (<50) businesses, BED	95%	94%
Ψ_H	Mass of "high" business opportunities	$9.2e-6$	Startup success rate, BED	21%	24%
$\underline{z}_H$	Mean "high" productivity	0.131	Average establishment size, BED	15.4	14.9
$\underline{z}_L$	Mean "low" productivity	0.104	Average establishment size of small (<50) businesses, BED	6.8	7.6
ρ	Productivity persistence	0.723	Establishment size life-cycle profile, BED	see Figure 12	
σ_z	Productivity shock dispersion	0.208	Establishment size life-cycle profile, BED	see Figure 12	
ζ_0	Capital adjustment costs, fixed	0.001	Establishment size life-cycle profile, BED	see Figure 12	
ζ_1	Capital adjustment costs, variable	0.61	Establishment size life-cycle profile, BED	see Figure 12	
δ	Exogenous exit rate	0	Establishment exit life-cycle profile, BED	see Figure 12	
μ_κ	Overhead costs, mean	0.795	Establishment exit life-cycle profile, BED	see Figure 12	
σ_κ	Overhead costs, dispersion	2.49	Establishment exit life-cycle profile, BED	see Figure 12	

Note: The table presents values for all model parameters. The first three columns describe each parameter and its parameterized value, while the last two columns indicate the most relevant target or data source.

of affording non-negative ("high") setup costs, $\kappa_H^\omega > 0$. Therefore, the fraction of large firms conducting remote work is informative about the magnitude of κ_H^ω.

Indirect inference: Remote work and firm outcomes. There are three key parameters determining how remote work choices impact firm outcomes: $\widetilde{f}$, $\widetilde{f}$ and $\widetilde{h}$. To pin these down, we will require our model to match key moments of remote work in the (pre-pandemic) U.S. economy.

In order to measure the extent of remote work, we follow Barrero et al. (2023) and focus on individuals' time allocated to work and its location as reported in the American Time Use Survey. Specifically, we count working days of individual i, d_i, as those in which individuals devote at least 6 hours to work in their main job.[15] Analogously, we define days worked from home, d_i^{home}, as those in which individuals spend at least 6 hours working from home in their main job. The *work from home rate*, ω_t, is then defined as the number of days spent working at home, d_t^{home} as a fraction of all work days, d_t. We do so at the sector level, s, and quarterly frequency:

$$\omega_{s,t} = \frac{\sum_{i=1}^{I_{s,t}} d_{i,\tau}^{home}}{\sum_{i=1}^{I_{s,t}} d_{i,\tau}}, \tag{82}$$

where $I_{s,t}$ is the number of individuals reporting in sector s in quarter t.

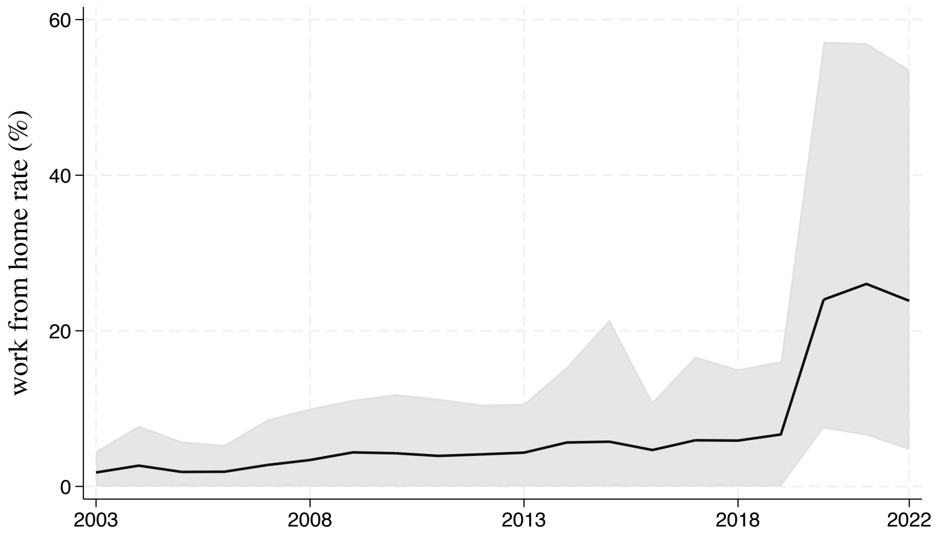
Figure 11 shows how work from home rates evolved over time. The solid black line depicts the *aggregate* work from home rate, while the shaded areas then indicate the range of work from home rates across industries (more details are presented in the Appendix). As is clear from the figure, work from home was very rare prior to the pandemic, albeit on an increasing trend. In particular, the average pre-pandemic remote work rate was 4%, increasing from about 1.8% in 2003 to 6.7% in 2019.

Note that with the help of ASEC and the CPS, we can obtain the same measure of remote work rates for different firm size bins. This reveals that on average remote work rates in large firms (with more than 100 workers) were 1.12 higher than those in all other businesses prior to the pandemic. In other words, larger businesses tended to conduct more of their production remotely.

Using the above information, we parameterize the key remote work parameters as follows. First, recall from Propositions 5 and 6 that all three factors – f, g and h –

[15]To define our baseline measure of working from home, we focus on workers with minimum real annual earnings of $20,000 (counted as 52 times average weekly earnings, deflated by the Personal Consumption Index). The Appendix shows that results are similar when considering "work-outside-workplace", i.e., anywhere but the respondent's workplace. Moreover, similar results are obtained when defining working from home as the fraction of hours worked from home *at the individual level*. Intuitively, this is because most individuals either spend entire days working at home or at their workplace.

Figure 11: Work from home rate



Note: The figure shows work from home rates – computed from ATUS as described in the main text – over time for the aggregate economy (solid black line) and the range of values across industries (shaded area).

impact firms' remote work choices. However, non-wage cost savings of remote work, g, induce heterogeneity in remote work rates across the firm size distribution. Therefore, to pin down f and g, we require our model to match the average remote work rate overall and among large firms. In the data, the average remote work rate was 4% prior to the pandemic and large firms had 12% higher remote work rates compared to small businesses.

The resulting parameterization implies that a business employing 4% of its workers remotely (the pre-pandemic average) would experience a 0.02% efficiency loss and a 1.3% lower non-wage cost. These results are broadly consistent with the empirical evidence that partial remote work arrangements come with little to no productivity or cost changes (see e.g., Barrero et al., 2023). Moreover, while detailed research in this area is still emerging, the few existing studies estimate that *fully remote* work is associated with productivity losses in the range of about $8-19\%$ (see Battiston et al., 2021; Yang et al., 2022; Emanuel and Harrington, 2023; Gibbs et al., 2023). Our baseline parameterization implies such costs to be about 14%, falling well within the empirical bounds.

Finally, while workers may have well preferred working from home even prior to the pandemic, we interpret the rare incidence of remote work prior to the pandemic as suggesting that working from home was not a significant factor in wage determination. Therefore, we set $\widetilde{h} = 0$ in the baseline (pre-pandemic) economy implying that firms' remote work decisions do not influence their wage bill.

Indirect inference: Permanent productivity differences. As with remote work setup costs, there are two types of firms when it comes to long-run productivity. To pin down the levels and respective masses of low- and high-productivity business opportunities, we target moments of the firm size distribution and the probability of starting up from the Business Employment Dynamics.

In particular, to discipline Ψ_L, Ψ_H, $\underline{z}_L$ and $\underline{z}_H$, we target the following four moments: (i) average size overall, (ii) average size of small firms (those with fewer than 50 workers), (iii) the share of small firms and (iv) the average startup probability ($M/(s_H + s_L)$). The first three targets are directly observable in the BED. To measure the last target, we interpret the model-implied startup probability as the *within first year* survival probability in the BED data.

Indirect inference: Non-wage labor costs. To parameterize non-wage labor costs, κ_n, we use information on firms' rental expenses from Compustat. Specifically, rental expenses in Compustat represent all costs for rental of land, space, buildings and/or equipment for continuing operations (see the Appendix for descriptive statistics). Through the lens of our model, the observed rental expenses are given by $g(\omega)\kappa_n n$. Therefore, to parameterize κ_n, we target the following reduced form relationship between rental expenses and employment in the cross-section of firms:

$$\ln(rent_{j,t}) = \alpha + \beta \ln n_{j,t} + \mathbf{X}'_{j,t}\theta + \epsilon_{j,t}, \tag{83}$$

where $rent_{j,t}$ are the rental expenses of firm j in period t and where the regression is estimated on pre-pandemic (2003-2019) data only. To account for potential differences across industries and for heterogeneous rental price developments across geographical locations, we include sector-year and city-year fixed effects in $\mathbf{X}_{j,t}$.

The coefficient of interest is estimated at $\beta = 0.861$ with a standard error of 0.0003. Using model-simulated data, and weights aligning the model-implied size distribution with that of Compustat as explained above, we replicate this reduced-form regression with $\kappa_n = 0.255$.

Indirect inference: Stochastic operational costs. Recall that firms in our model face an exogenous probability of shutting down, δ, as well as the option to exit the market endogenously. The latter occurs if the realization of stochastic fixed operational costs is high enough.

Therefore, to discipline the exogenous exit rate, the mean (μ_κ) and dispersion (σ_κ) of stochastic operation costs, we target the entire life-cycle profile of exit rates from age 1 to age 20 taken from the BED and averaged over the years 2003 to 2019.[16] Intuitively, the level of exit rates among young and old firms is informative about δ and μ_κ. Moreover, the speed at which exit rates decline with age is informative about σ_z. The right panel of Figure 12 shows the empirical and model-implied exit rates.

Indirect inference: Firm-level productivity and capital adjustment costs. The remaining parameters pertain to firm-level productivity shocks (persistence, ρ, and dispersion, σ_z) and to capital adjustment costs (ζ_0 and ζ_1). All four of these parameters are closely related to firm size and growth patterns.

Therefore, to pin these parameters down, we require our model to match the entire life-cycle profile of business size from startup (age 0) to age 20 as observed in the BED. The left panel of Figure 12 shows the empirical and model-implied size profiles.

4.3 Model Performance

Table 4 and Figure 12 show the targeted moments and their model counterparts. In addition, our model is consistent with a range of *untargeted* moments and estimates in the literature.

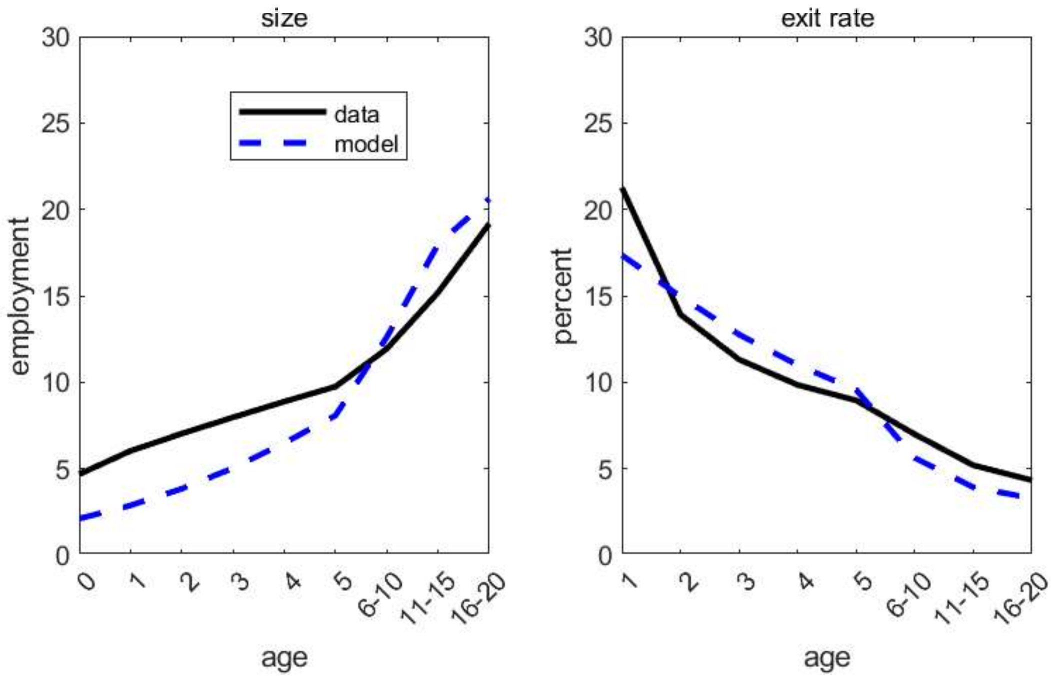
Investment, firm-level productivity and the firm-size distribution. First, our model is consistent with capital investment patterns. In particular, Cooper and Haltiwanger (2006) estimate average investment rates at around 12% and average inaction rates (investment rates between -1% and 1%) of about 8%. Our model predicts these values to be, respectively, 14% and 7%.

Second, in addition to matching average patterns, our model also does well in matching dispersion moments. Specifically, Cooper and Haltiwanger (2006) report the dispersion of investment rates to be 0.34. Our model predicts this value to be 0.36.

Third, the implied values of persistence and volatility of firm-specific productivity are close to empirical estimates in existing studies. For instance, Foster, Haltiwanger and Syverson (2008) estimate persistence of firm-specific TFP to lie between 0.75 and 0.81. The standard deviation of such productivity shocks is then estimated to fall within the range of 0.21 and 0.26. Our parameterization strategy yields a persistence parameter of 0.72 and a standard deviation of productivity shocks of 0.21. In addition, the implied firm-level growth process is also consistent with the evidence on high-growth firms. In particular, the share of gazelles – businesses with growth rates exceeding 25% – is about

[16] Since the BED starts in 1992, the life-cycle information for establishments in the age group of 6-10 years is from 2004 to 2019, for ages 11-15 is from 2009 to 2019, and for the age group 16-20 it is from 2014 to 2019.

Figure 12: Life-cycle profiles of size and exit rates: Data and model



Note: The left panel shows average establishment size (employment) by age, while the right panel shows average exit rates by age for the model and the data. The latter is taken from the BED, averaged over the years 2003-2019.

9 percent, consistent with the U.S. data (see Haltiwanger, Jarmin, Kulick and Miranda, 2016).

Fourth, as will become clear, the share of small firms will be important for our quantitative results. While our model is designed to match the share of small businesses (< 50 workers), it also does well at matching the share of very small businesses (with 1-4 workers). This is true both overall and among startups only. In particular, while in the data the share of very small establishments among all businesses (among startups) is 0.54 (0.89), in the model this share is 0.55 (0.89).

Additional work from home patterns. While we target the overall share of firms doing remote work and that among large firms (with more than 100 employees), our model also matches the extensive margin of work from home in other parts of the firm size distribution. In particular, while 24 percent of businesses with fewer than 20 workers report doing remote work in the data, this fraction is 23 percent in our model. On the other extreme, 44 percent of very large firms (with more than 500 workers) have some of their employees work remotely in the data. In our model, this fraction is also 44 percent.

Finally, let us describe in detail the model-implied remote work patterns, summarized in 5. The table reports remote work rates "unconditionally" for all firms and "conditionally" for businesses conducting remote work. In addition, we report work from home

Table 5: Remote work rates in the model (%)

	Size-weighted		Un-weighted			
	Uncond.	Cond.	Uncond.		Cond.	
	Mean	Mean	Mean	Std	Mean	Std
All	3.8	14.0	4.8	9.5	20.6	8.1
Firms with < 50 workers	3.7	16.0	4.9	9.8	21.4	8.1
Firms younger than 6 years	3.4	16.9	4.9	10.7	24.8	8.9
High-type firms	3.9	13.6	4.5	9.0	19.4	7.7
Low-type firms	3.7	14.8	5.0	10.0	21.7	8.3

Note: The first two columns of the table report size- (employment-) weighted means of remote work rates. The remaining columns compute unweighted means and standard deviations, all reported unconditionally ("Uncond.") and conditionally ("Cond.") on businesses conducting remote work. The rows indicate different firm groups: "all" firms , firms with less than 20 workers, businesses younger than 6 years, "high-type" firms (with $\underline{z}_H$), "low-type" businesses (with $\underline{z}_L$).

rates for various firm groups and for a size- (employment-) and un-weighted sample.[17] While the employment-weighted sample corresponds to the information in the ATUS-ASEC data (which is worker-based), to the best of our knowledge there is no dataset for the U.S. economy allowing to compute work from home rates at the firm-level. Therefore, one of the contributions of this paper is to use our model to provide such firm-based statistics.

Several patterns stand out. First, there is a large amount of heterogeneity in remote work rates (last and third-last columns). This holds true even within subgroups of firms. Second, conditional remote work rates are considerably higher than unconditional ones reflecting that – on average – only about 23 percent of businesses conduct remote work. Third, conditional and unconditional remote work rates are farther apart among smaller and younger firms. This is because such businesses are less likely to pay the setup costs of remote work. Fourth, conditional on conducting remote work, smaller and younger firms do more of it. Noting that young firms have only about 8 workers on average, this pattern reflects the fact that smaller businesses benefit relatively more from remote work. Finally, since low-type firms are on average smaller, their remote work patterns are similar to those of small businesses. Nevertheless, since high-type firms grow towards their larger size only gradually, their remote work rates are not dramatically different to those of low-type businesses.

[17] In computing these statistics, we exclude "non-employers" which we interpret as firms with employment below or equal to 1.

5 Macroeconomic Impact of the Remote Work Revolution

In this section, we use our model to quantitatively evaluate how changes in work from home patterns impact business dynamism and, in turn, the macroeconomy. Towards this end, we take our generalized model and compare it to a "remote economy", which is identical to our baseline but features higher remote work rates. The difference in model outcomes between the baseline and the remote economy then offers a quantification of the impact that more prevalent remote work arrangements have on the economy.[18]

5.1 The remote work revolution

As explained in the previous section, in our framework there are three key drivers of remote work decisions: $\widetilde{f}$, $\widetilde{g}$ and $\widetilde{h}$.[19] To discipline changes in these parameters, we adopt the same parameterization strategy as in the previous section and jointly target moments related to remote work which we describe in detail below.

Productivity and price of remote work. Recall that our baseline parameterization uses information on remote work rates overall and by firm size to pin down $\widetilde{f}$ and $\widetilde{g}$. Therefore, we will use exactly the same targets in the *post-pandemic* period to discipline the change in these parameters.

In particular, according to our ATUS-ASEC data, remote work rates increased substantially on average (see Figure 11). Specifically, the average remote work rate shot up from the pre-pandemic 4 percent to a post-COVID average of 24 percent. At the same time, the remote work rate among large firms (with more than 100 workers) relative to small businesses also increased. While before COVID this ratio was 1.12, it increased to 1.23 in the post-pandemic U.S. economy.

Matching these moments results in a parameterization which implies that remote work

[18]Note that when comparing the baseline to the remote economy, we consider their respective stationary steady states and ignore transition dynamics. We also note that while existing research points to reasons "Why working from home will stick" Barrero et al. (2021), it remains an open question to what extent the post-pandemic remote work rates will remain elevated. For instance, in September 2024 Amazon famously announced that it will go back to on-site production from 2025. Through the lens of our model, Amazon could simply be viewed as a business with high remote work setup costs. More importantly, however, and as will become clear below, a key force behind our model results is how more favorable or desired remote work increases the incentives of potential startups to enter the economy. At the end of this section, we provide evidence on startups consistent with our model predictions.

[19]We do not consider changes in setup costs κ^{ω}. In our view, the costs of setting up remote work did not change fundamentally. For example, to this date Zoom – the telecommunications platform offering virtual conferencing services – still offers its "Basic" plan free of charge.

becomes more productive ($\widetilde{f}$ increases from -0.151 to -0.128) and cheaper ($\widetilde{g}$ increases from 0.325 to 0.488). These changes are consistent with existing evidence that firms and workers are better positioned to work from home more effectively (see e.g., Barrero et al., 2022). That said, the quantitative implications for individual firms remain relatively modest. In particular, an average firm in the baseline economy would see its productivity losses decline from 0.024 to 0.021 percent and its cost savings increase from 1.3 to 1.9 percent.

Preferences for remote work. Recall that our baseline, pre-pandemic, economy assumed that (even if workers preferred remote work) the rare incidence of working from home did not influence wage determination. In contrast, existing evidence suggests that the COVID-19 pandemic cast work from home into the mainstream which, in turn, allowed workers to negotiate more freely for this new amenity value of jobs.

To allow for such forces, we release $\widetilde{h}$ in the remote economy and allow firms' wages to respond to remote work decisions. In order to discipline the magnitude of this change, we lean on existing empirical evidence about the value workers place on flexible work arrangements.

For example, Mas and Pallais (2017) provide experimental evidence that workers are on average willing to sacrifice 8% of wages in return for a hybrid work setup. Using survey evidence, Vij, Souza, Barrie, Anilan, Sarmiento and Washington (2023) estimate that an average worker in Australia is willing to forgo $4-8\%$ of wages for the possibility of remote work, while Barrero et al. (2021) document that U.S. businesses can reduce wage growth pressures by about 1% with a roll-out of remote work. In related studies, Bloom et al. (2015) and Angelici and Profeta (2023) find positive effects of remote work on workers' job satisfaction and Autor, Dube and McGrew (2024) suggest that changes in the amenity-value of jobs played a role in the "unexpected compression" of the age distribution following the pandemic.

To parameterize $\widetilde{h}$ in the remote economy, we use the mid-point of the above values. In particular, we require that in our remote economy workers are willing to sacrifice about 4% of wages in return for a hybrid work arrangement, i.e. $h(0.5) \approx 0.96$.

5.2 Drivers of the Remote Work Revolution

As a first step in our analysis, we use our framework to quantify the underlying drivers of the remote work revolution. To do so, we consider three versions of the remote economy.

In particular, we hold one of the three channels ($\widetilde{f}$, $\widetilde{g}$ or $\widetilde{h}$) unchanged at its baseline value, while letting the other two channels take on values of the remote economy described in Section 5.1. For example, to gauge the impact of changes in the productivity of remote work we compare our remote economy to a counterfactual in which $\widetilde{f}$ is held at it's baseline value, while $\widetilde{g}$ and $\widetilde{h}$ are assumed to take on the remote economy values. Comparing these three versions to the remote economy (in which all three channels change) isolates the importance of the omitted factor. Three key messages stand out.

Preferences for remote work. First, while all three changes (improvements in productivity, declines in costs and stronger preferences for remote work) lead firms to employ a larger share of their workforce remotely, changes in preference are the dominant driver. In particular, had it not been for workers' stronger desire to work from home, the average remote work rate would be "only" 11 percent, instead of the observed 24 percent. In other words, almost 2/3 of the surge in remote work rates is driven by a change in workers' preferences.

This finding is consistent with recent evidence. For instance, Zarate et al. (2024) find that cultural differences and "individualism" accounts for about 1/3 of the cross-country variation in remote work rates. Similarly, using a calibrated labor market model accommodating amenity-value of jobs, Bagga et al. (2024) find a crucial role of changes in workers' preferences in explaining labor market patterns following the pandemic, including the rise in remote work. Finally, using survey evidence Barrero et al. (2023) find that workers' preferences for a (hybrid) remote work setup are "remarkably prevalent" since the pandemic.

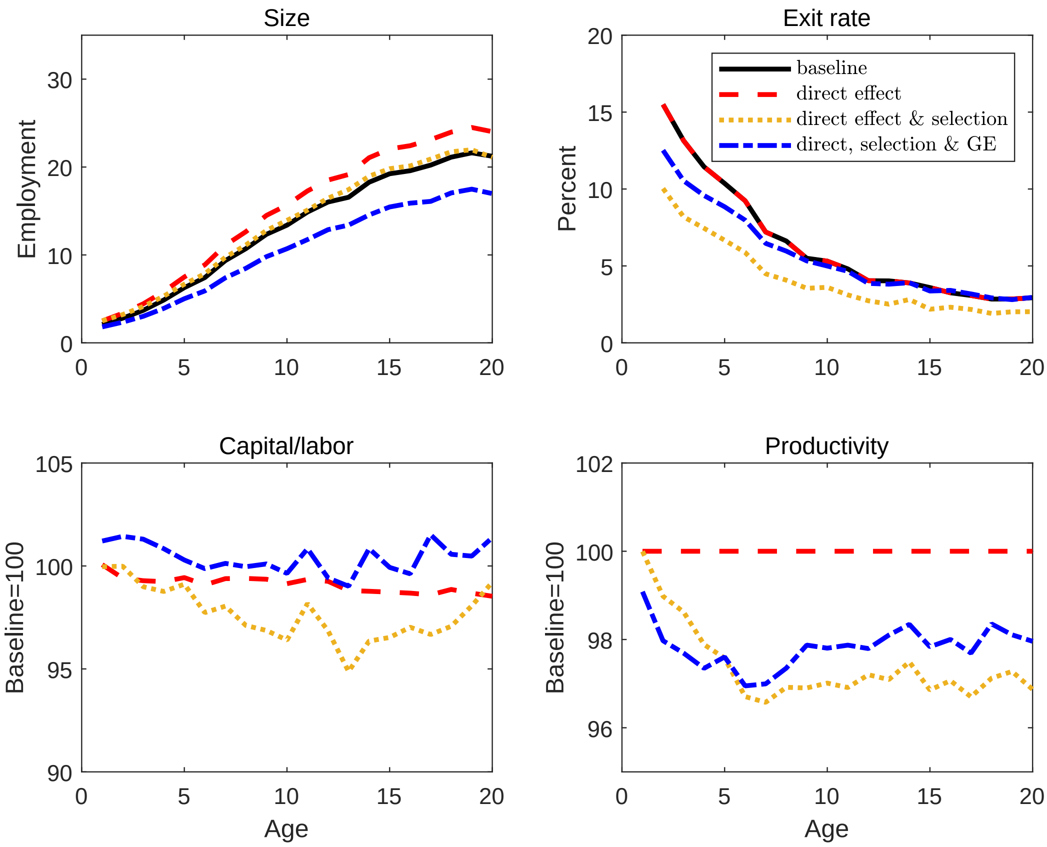
Productivity and cost of remote work. Second, changes in the efficiency of remote work are, intuitively, most important for average firm-level productivity. In particular, had it not been for remote work becoming more efficient, average firm-level productivity would have been almost 1 percent lower compared to the remote economy.

Third, lower costs of remote work are crucial for a shift in the distribution of firms towards smaller businesses. This is intuitive in light of our Proposition 5 which highlights that especially small firms benefit from reductions in (fixed) costs associated with remote work. As will become clear below, this model prediction is also key for understanding the macroeconomic impact of the remote work revolution. We provide detailed empirical evidence consistent with this finding in the following section.

5.3 Remote Work and Business Dynamism

In this subsection, we quantify the connection between work from home decisions and business dynamism. To isolate changes in firms' choices from shifts in the composition of businesses, we first separately discuss firms which "always" and "never" conduct work

Figure 13: Higher remote work rates: Effects on firms which always conduct remote work



Note: The figure shows average firm-level employment (top left panel), exit rates (top right panel), capital-to-labor ratios (bottom left) and productivity (bottom right panel) as a function of firm age. It does so for the "baseline" model, and for the case when remote work is cheaper and more efficient. The latter is shown in partial equilibrium, ignoring firm selection and GE effects ("direct effect"), in partial equilibrium with firm selection ("direct effect & selection") and in the new general equilibrium ("direct effect, selection & GE"). The bottom two panels are expressed relative to values in the baseline model. All panels are for firms which always conduct remote work from startup, both in the baseline and in the remote economy.

from home. Thereafter, we turn towards changes in the composition of firms and to the overall impact on business dynamism.

Firm growth and selection: Firms which always conduct remote work. Figure 13 displays how the remote work revolution affects average firm-level employment (top left panel), exit rates (top right panel), capital-labor ratios (bottom left) and productivity (bottom right panel). Each of these are plotted over the life-cycle of firms which always choose to pay the remote work setup cost at entry – both in the baseline and the remote economy.

In addition to our baseline specification (black solid line), we consider 3 different scenarios of the remote economy. First, a partial equilibrium response which ignores both firm selection effects (entry and exit) and changes in equilibrium prices – this is shown by the "direct effects" line. Second, we consider the same partial equilibrium response, but allow for firm selection (changes in entry and exit), while keeping wages fixed – this is shown by the "direct effects & selection" line. Finally, we also plot the impact in general equilibrium (GE) allowing for a change in wages – this is shown in the "direct effects, selection & GE" line. The latter corresponds to the final stationary steady state of our remote economy.

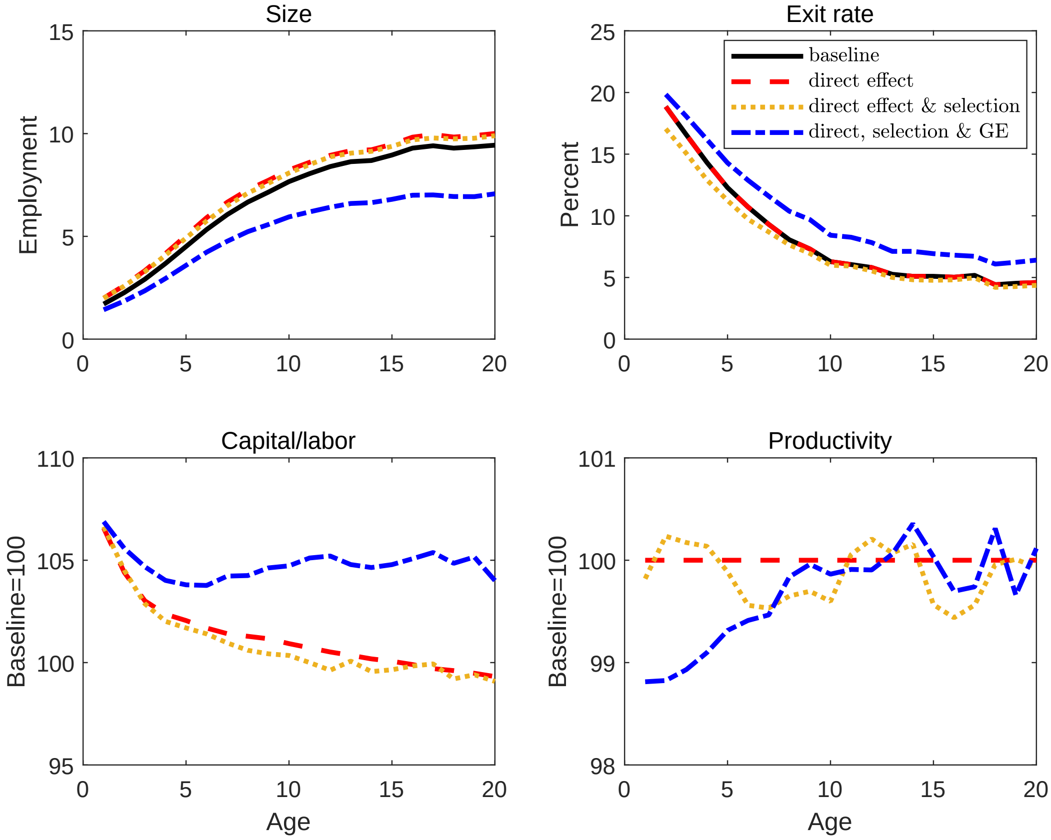
First, ignoring firm selection and general equilibrium effects, firms decide to expand production (top left panel) when remote work becomes cheaper and more efficient. In doing so, firms slightly reduce their capital-labor ratios as they take advantage of the relatively cheaper production factor (bottom left panel). By construction, average TFP (which excludes efficiency losses of remote work, $f(\omega)$) and exit rates are unchanged when ignoring selection and GE effects (right panels).

Next, as predicted by our theoretical analysis in Section 2, more favorable remote work conditions raise profits and firm values which induce greater entry and reduce firm exit (top right panel). Note that firm exit declines more for younger firms. This happens because younger firms are on average smaller and for such businesses the reduction in (fixed) costs related to work from home is relatively more beneficial. Therefore, some firms which could not afford to stay in business when remote work was costlier can now remain in operation. This selection effect pulls down average firm productivity (bottom right panel) and with it also average firm size (top left panel). We will return to this effect when evaluating the macroeconomic impact of more prevalent remote work.

Finally, with increased entry and lower exit, the number of firms expands. This raises labor demand and with it the equilibrium wage. Such higher labor costs induce firms to scale down production (top left panel) and shift towards capital as a production factor, increasing capital-to-labor ratios above those in the baseline economy (bottom left panel). In addition, higher production costs make it harder for all businesses to survive and exit rates increase across the board – though less so for small firms (top right panel). Because of such weaker firm selection among young firms, average firm-level productivity remains below the baseline. While higher exit rates among older firms lead to a partial productivity catch up, there remains a persistent productivity gap between the remote and baseline economy (bottom right panel).

Firm growth and selection: Firms which never conduct remote work. Figure 14 turns towards firms which never conduct remote work – neither in the baseline, nor in the remote economy. Ignoring selection and general equilibrium effects, cheaper and more efficient work from home leads to an expansion of production even among businesses which operate on-site only (top left panel). The reason for this is that as remote work becomes more favorable, a larger fraction of businesses *expect* they may choose to pay

Figure 14: Higher remote work rates: Effects on firms which never conduct remote work



Note: The figure shows average firm-level employment (top left panel), exit rates (top right panel), capital-to-labor ratios (bottom left) and productivity (bottom right panel) as a function of firm age. It does so for the "baseline" model, and for the case when remote work is cheaper and more efficient. The latter is shown in partial equilibrium, ignoring firm selection and GE effects ("direct effect"), in partial equilibrium with firm selection ("direct effect & selection") and in the new general equilibrium ("direct effect, selection & GE"). The bottom two panels are expressed relative to values in the baseline model. All panels are for firms which never conduct remote work from startup, both in the baseline and in the remote economy.

the setup cost at some point in the future. This, in turn, makes businesses front load the costs of building up capital in expectation of being able to take advantage of cheaper labor in the case of going remote (bottom left panel).

The more favorable continuation values, as well as a larger capital stock, reduce firm exit rates slightly (top right panel). Quantitatively, however, this impact is very small and the effect on average TFP is negligible (bottom right panel). However, higher equilibrium wages result in a strong decline in firm size and a rise in exit rates (top panels). Therefore, while fully on-site firms benefit from cheaper and more efficient remote work only indirectly (in expectation), they are directly negatively affected by the general equilibrium increase in wages for which they are not "responsible".

Composition of firms. Let us now turn to investigating how the composition of firms differs between the baseline and the remote economy. First, the share of firms deciding to conduct remote work is higher in the remote economy, at about 40%. More importantly, however, the composition of firm types is different since low-productivity firms (which are on average smaller) benefit relatively more from cheaper work from home. In particular, the share of high-type firms entering the economy drops by more than 13 percent. Moreover, there is also a shift in exit rates with high-type firms seeing their survival rates increase *relatively* more compared to those of low-type firms.[20]

Overall, there is a clear pattern of "winners" and "losers" from the remote work revolution. The winners are small (on average low-productivity) businesses conducting remote work. The losers are larger (typically high-productivity) businesses with high remote work setup costs. While these businesses cannot take advantage of the more favorable work from home conditions, they do feel the pain of the higher equilibrium wage. Quantitatively, compared to the baseline, low-productivity and low-setup cost firm types are almost 50% more common in the remote economy (a firm share of 18.5% vs 12.9%). In contrast, high-productivity and high-setup cost firm types are almost 20% less common in the counterfactual (a firm share of 27.9% vs 33.8%).

5.4 Remote Work and the Macroeconomy

Intuitively, cheaper, more efficient and more desirable remote work frees up existing resources (and creates new ones) and leads to an economic expansion. However, the changes in business dynamism described above, and in particular the changes in the (productivity) composition of firms, serve to offset some of the positive impacts of the remote work revolution. In this section, we quantify these counteracting forces at the macroeconomic level.

Aggregate output and TFP. Let us begin with defining "aggregate TFP" which in our framework is given by:

$$Z = \frac{Y}{(N^{\alpha}K^{1-\alpha})^{\theta}} = \sum_i \sum_l \int \int z f(\omega) \left(\left(\frac{n}{\overline{n}}\right)^{\alpha} \left(\frac{k}{\overline{k}}\right)^{1-\alpha} \right)^{\theta} \Omega^{1-\theta} \widetilde{\mu}_{i,l}(z,k) dz dk, \qquad (84)$$

where we will refer to the term $(N^{\alpha}K^{1-\alpha})^{\theta}$ as the "scale" of the economy and where bars indicate averages, such that $N = \overline{n}\Omega$ and $K = \overline{k}\Omega$, and where $\widetilde{\mu} = \mu/\Omega$ is the probability distribution function.

The expression above highlights four drivers of aggregate TFP. First, the distribution of firms across (long-run) productivity levels, $\widetilde{\mu}$. Recall that this is a combination of

[20] Note that the *level* of exit rates among high-productivity firms remains considerably lower compared to those of low-productivity businesses. This is true both in the baseline and remote equilibrium.

startup composition and survival rates – both of which are endogenous in our framework. Second, endogenous remote work choices which impact firm-level efficiency, $f(\omega)$. Third, the allocation of inputs across heterogeneous firms, $\left((n/\overline{n})^{\alpha}(k/\overline{k})^{1-\alpha}\right)^{\theta}$. Fourth, the mass of firms, $\Omega^{1-\theta}$, since a greater mass of smaller businesses improves efficiency in the presence of decreasing returns to scale.

Starting with Panel A of Table 6, the first row shows that aggregate output is about 2.9 percent higher in the remote economy. This is the result of both a slightly higher aggregate productivity (second row, first column) and an increased scale of the economy (second row, second column). However, the single most important contributor to both of these is the higher mass of firms, Ω, in the remote economy. In fact, as we have highlighted before, the distribution of firms, $\widetilde{\mu}$, shifts towards low-productivity firm types, dragging down aggregate TFP. We will return to this point below.

Consumption and welfare. The next three columns of Table 6 show how consumption differs between the remote and baseline economies and splits this gap into the contributions of output, investment and costs ($C = Y - I - Costs$). The latter encompass capital adjustment costs, fixed operation costs, non-wage on-site costs, remote work setup costs and the costs of entry.

Overall, Panel A shows that consumption is considerably higher in the counterfactual economy, predominantly driven by a rise in aggregate output.[21] In contrast, higher investment (consistent with the increased aggregate capital stock) dampens consumption. This effect is roughly offset by a decline in paid costs, as spending on overhead costs and non-wage labor costs falls due to less costly and more prevalent remote work.

Finally, the last two columns of Table 6 report differences in household welfare ($\mathcal{W} = \log(C) - \sum_j v_j n_j$). Our model predicts that welfare is slightly higher in the remote economy. This is entirely driven by the strong consumption increase. In contrast, households' optimal labor supply is somewhat higher in the remote economy (though accompanied with a lower disutility due to increased remote work rates). Quantitatively, however, this latter effect does not overturn the welfare benefits of higher consumption.

5.5 More firms vs their composition

The remote economy enjoys higher productivity, more output, consumption and increased welfare. In this subsection, we highlight the crucial role of firm entry for these aggregate

[21] The reason why the percentage contribution of output towards consumption is higher than what is reported in the first two columns is that it takes into account the output share in consumption which – given that investment and costs enter negatively – is higher than one. A similar effect holds for the contribution of consumption to welfare.

Table 6: Impact of increased remote work: Changes in aggregates (in %)

Panel A: Full adjustment

	Output (Y)		Consumption (C)			Welfare ($\mathcal{W}$)	
Overall	2.9		4.3			0.2	
Components	Z	$(N^\alpha K^{1-\alpha})^\theta$	Y	I	Costs	C	N
	1.1	1.7	4.3	−0.8	1.0	0.3	−0.1

Panel B: No change in the mass of firms

	Output (Y)		Consumption (C)			Welfare ($\mathcal{W}$)	
Overall	−1.5		−0.1			0.0	
Components	Z	$(N^\alpha K^{1-\alpha})^\theta$	Y	I	Costs	C	N
	−0.6	−1.0	−2.3	0.2	2.0	−0.0	0.0

Note: The first row in each panel of the table shows log-changes in aggregate Output (Y), Consumption (C) and Welfare ($\mathcal{W}$). The second row then split the overall changes into the contributions of the various components. All values are reported in percent differences from the baseline economy. Panel A shows the "full adjustment", Panel B considers a scenario with "no change in the mass of firms".

gains of the remote work revolution.

Remote economy with barriers to firm entry. To make our point, we consider a variant of the remote economy in which firm entry is subdued. We do this as a reduced form way of modeling frictions (e.g., financial or regulatory) impeding a flexible entry response. Practically, we control the extent of firm entry by replacing the entry function (75) with an exogenous rule such that the remote economy features exactly the same mass of firms as the baseline model. In doing so, however, we keep the rest of the model exactly the same, still match all the same moments as before and solve for the general equilibrium wage. The results are in Panel B of Table 6 and the details of our computational strategy are in the Appendix.

Remote work revolution without a surge in startups. Panel B of Table 6 shows the aggregate impact of the remote work revolution in the case where firm entry is held constant. In this case, aggregate TFP is *lower* in the remote economy compared to the baseline. This happens despite the fact that remote work is *more* efficient (i.e., $f(\omega)$ improves) in the remote economy.

The reason lies in the different business dynamism patterns in the remote economy, discussed in Section 5.3. In particular, the remote economy shifts towards smaller, less productive, businesses – the "winners" of more favorable remote work. The losers, typically high-productivity businesses with high remote work setup costs, are effectively crowded out through the general equilibrium forces.

As a result, aggregate output is about 1.5 percent lower in the counterfactual economy. While lower investment and paid costs help offset some of this drop, aggregate consumption still declines slightly. In equilibrium, households choose to reduce their labor supply and, as a result, welfare is effectively the same in both economies.

Therefore, while cheaper, more efficient and more desirable remote work arrangements are, all else equal, unambiguously positive for individual workers and firms, they also induce changes in business dynamism. In the absence of increased firm entry and the associated economic boom, the shift towards less productive firms lowers aggregate productivity and leaves welfare effectively unchanged. Therefore, a key insight from our framework is that the overall aggregate effects of the remote work revolution crucially rest on the ease with which startups can enter the economy. We return to this point at the end of the paper when discussing cross-country evidence.

6 Model Validation and Discussion

In this section, we offer empirical support for key model predictions and underlying channels. First, we show that the model is consistent – qualitatively and quantitatively – with the post-pandemic evolution of the U.S. macroeconomy. Next, using micro-data and a novel identification strategy, we provide empirical evidence in support of the key model trade-off between an entry boom and shifts in the composition of firms. These two sets of exercises, therefore, further help validate our theoretical framework.

6.1 Model Predictions in the Data

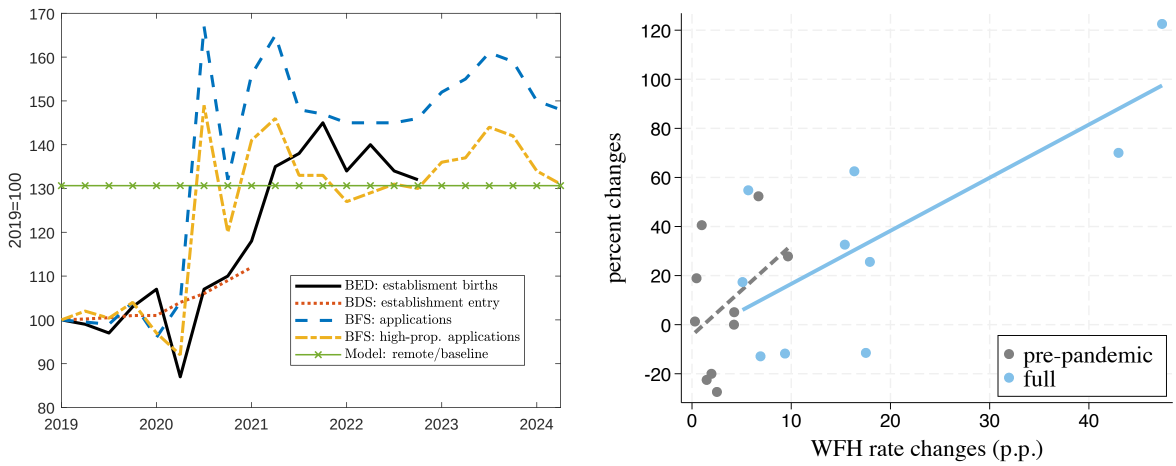
As a first step, we confront key model predictions about changes in business dynamism with those observed in U.S. data. In doing so, we focus on the fundamental trade-off between firm entry and changes in the composition of businesses.

Firm entry. As explained in the previous section, a crucial force behind the positive aggregate effects of the remote work revolution is firm entry. We now show that the extent of the model-implied increase in entry is in fact quantitatively reasonable.

Specifically, the U.S. currently finds itself in a "surprising surge in applications for new businesses" (see Decker and Haltiwanger, 2024, p.1). The left panel of Figure 15 shows various recent measures of (planned) business entry in the U.S. economy. These include actual establishment entry taken from the BED and BDS datasets, as well as business applications from the Business Formation Statistics (BFS) of the Census Bureau. The figure displays both overall applications as well as "high-propensity" applications which

Figure 15: Model validation: Entry

(a) Aggregate level (b) Industry level



Note: The left panel of the figure shows recent measures of business entry: establishment births from the BED, establishment entry from the BDS, overall applications from the BFS and "high-propensity" applications from the BFS. BFS data are quarterly averages of monthly series, while the BDS data is annual but interpolated to a quarterly frequency. In addition, the left panel also shows the model-implied mass of entrants in the remote economy relative to that in the baseline ("Model: remote/baseline"). The right panel shows changes in remote work rates and entry at the industry-level. Work from home rates are estimated from ATUS as described in the main text. Business entry is taken from the BED. The right panel shows data for the pre-pandemic sample (2003-2019) and the full sample (2003-2022).

are deemed as likely converting into actual entry and employment. The latter two are the most timely as the BFS statistics are published monthly. In contrast, the BDS are annual and published with the longest lag.

The figure shows that all measures picked up strongly in 2020, with the BED and BFS measures reaching a new, higher plateau since about 2022. The latest evidence seems to point towards a sustained entry rise. Through the lens of our model, this is consistent with persistently more favorable remote work conditions. Importantly, the figure also shows that the extent of the model-implied rise in firm entry in the remote economy relative to the baseline ("Model: remote/baseline") is quantitatively reasonable, lying on the conservative end of the empirical measures.

While the left panel of Figure 15 focuses on the mass of entrants, the first column of Table 7 reports the change in the entry rate (startups as a share of all businesses). Also in this dimension, our model performs well, accounting for about 46 percent of post-pandemic the entry rate increase observed in the data.[22]

The right panel of Figure 15 further documents that the positive relationship between increases in remote work and firm entry is not limited to the pandemic period or the

[22] The difference between the mass of startups and the entry rate is driven by a stronger model-implied increase in the total number of firms relative to the data. This is to be expected since the model compares long-run equilibria, while in the data the mass of firms is likely still in transition.

aggregate economy. In particular, the horizontal axis shows percentage point changes in industry-level work from home rates, while the vertical axis depicts the corresponding percent changes in the numbers of entrants. In all these cases, we consider separately the full sample (2003-2022) and the pre-pandemic period (2003-2019).[23] The figure shows that – consistent with our theoretical model – increases in work from home rates are clearly associated with strong increases in firm entry.

Composition of firms. While a boom in firm entry drives the aggregate benefits of the remote work revolution, in our model these effects are dampened by a shift towards smaller (less productive) firms. Figure 16 and Table 7 show that this prediction is also qualitatively and quantitatively consistent with the data.

Specifically, the left panel of Figure 16 shows that increases in remote work rates are on average associated with declines in firm size. This holds both before and after the pandemic. This prediction is also confirmed for the aggregate economy in Table 7. In particular, the last four columns of the table show the percent changes in firm sizes of various groups of businesses in the data and the model. Aside from an overall decline in firm size, our framework also aligns well with the observed heterogeneity in firm size declines between firms of different ages and between entering and exiting businesses.

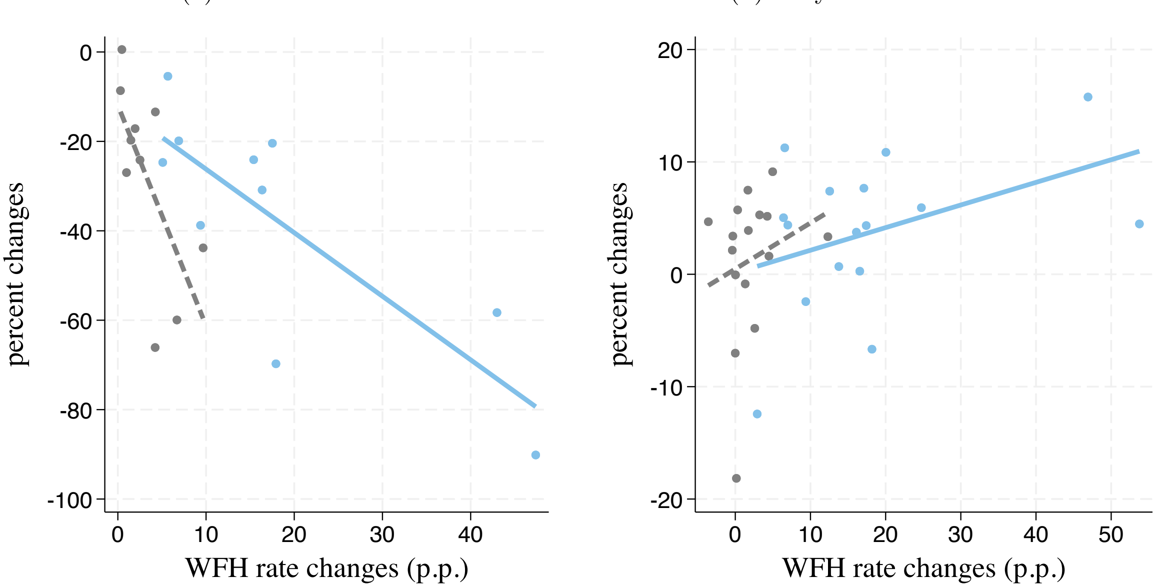
Next, the right panel of Figure 16 documents that increases in remote work are associated with shifts in the entire firm size distribution towards (very) small businesses. Moreover, in the aggregate, the share of very small establishments (1-4 workers) increased by about 4 (5) percentage points among entrants (all businesses). In comparison, our model predicts these shifts to be 5 (7) percentage points, respectively. Therefore, the key trade-off between increased entry and a shift towards small firms highlighted by our framework is also present in the data.

Firm exit. Before moving on, let us highlight that our model is also consistent with the observed increase in firm exit rates. Moreover, it correctly predicts that this increase is mainly concentrated in older businesses. Through the lens of our model, this is because younger businesses tend to be smaller and, therefore, benefit relatively more from the remote work revolution.

Overall, the above evidence suggests that the key model predictions are observed in the data. Moreover, the strength with which business dynamism changes – both in terms

[23] We compute changes as the difference in the respective values at the end of the pre-pandemic period (average of 2018Q1-2019Q4) or the full sample (average of 2021Q1-2022Q4) relative to the start of our sample (average of 2003Q1-2004Q4). More details are in the Appendix, including panel regressions with additional controls which confirm our results in Figure 15.

Figure 16: Model validation: Entrant size



Note: The figure depicts super-sector changes in work from home (WFH) rates on the horizontal axis (in percentage points) and (percent) changes in average entrant size (Panel a) and the fraction of very small (1-4 worker) firms (Panel b). Work from home rates are estimated from ATUS as described in the main text. Business sizes are taken from the BED. All panels show data for the pre-pandemic sample (2003-2019) and the full sample (2003-2022).

Table 7: Changes in business dynamism (in %)

		Exit rate			Size			
	Entry rate	All	Young	Old	Entrants	Young	Old	Exiters
Data	+28	+15	+1	+22	−24	−18	−15	−22
Model	+13	+13	+5	+19	−16	−18	−15	−21

Note: The table shows changes in the "entry rate", "exit rate" (of young, old and all firms), and firm "size" (of entrants, young, old and exiting firms). Young is defined as less than 6 years, "old" refers to 16-20 year old firms. The top row shows the BED data, while the bottom row shows model-predicted differences based on a comparison of the remote and baseline economies.

of the firm size distribution and aggregate business entry and exit – is also in line with the empirical evidence.

6.2 Model Mechanism in the Data

As a final step in our analysis, we use firm-level data to provide further evidence in support of the underlying mechanisms operating in our model. In particular, based on existing research (see Section 4), our framework assumes that at the firm-level remote work reduces (i) productivity, (ii) wage costs and (iii) non-wage costs (both fixed and variable).

A key empirical challenge for an independent validation of these mechanisms is that, to the best of our knowledge, there is no comprehensive *firm-level* information on remote work rates in the U.S. economy. Therefore, to identify the impact of changes in remote work on firms' costs and productivity, we propose a novel identification strategy.

Impact of remote work on firm-level outcomes: Identification. To gauge the impact of changes in remote work conditions on firm-level outcomes, we propose a novel identification strategy. In particular, we will exploit information on firms' "rental commitments" as reported in Compustat. Concretely, rental commitments represent the minimum rental expense due in the next five years reported in firms' balance sheets under all existing non-cancelable leases.

Importantly, rental commitments are reported separately for up to five years ahead. Therefore, to the extent that firms in 2019 did not predict the need for flexibility in their rental contracts (to help them deal with the pandemic-induced forced work from home), the information on rental commitments constitutes exogenous variation in the *exposure* of firms to the remote work revolution.

Intuitively, businesses with no future rental commitments were "exposed" to the remote work revolution because they were free to adjust their rental expenses. In contrast, businesses with rental commitments up to five years in advance were not exposed. Even if their employees started to work remotely during the pandemic, their rental costs were pre-committed. Moreover, because exposed firms can reap the cost-saving benefits of remote work, they will be relatively more incentivized to choose higher remote work rates. Therefore, while our rental commitment exposure measure is most closely related to rental costs, we will also use it in the context of other firm-level variables. Most notably, firm-level productivity, wage costs and fixed costs – the channels present in our model.

In the data, about 40 percent of all firms report both rental expenditures and commitments. Among these, firms are roughly equally split between no commitments and rental commitments five years into the future. Moreover, looking into the history of these firms, there is little evidence of "commitment types." For instance, out of the firms reporting commitments in 2019, more than 2/3 reported having no commitments at some point in the previous 15 years. Therefore, these statistics suggest that rental commitments in 2019 represent exogenous variation across firms.

Impact of remote work on firm-level outcomes: Estimation. To gauge the impact of the work from home revolution on firm-level outcomes, we use Compustat data between 2016 and 2023 to estimate the following set of regressions:

$$\ln y_{j,t} = \alpha + \beta_E \mathbb{1}(E)_j + \beta_T \mathbb{1}(T)_{j,t} + \beta_{E,T} \mathbb{1}(T)_{j,t} \mathbb{1}(E)_j + \mathbf{X}'\theta + \epsilon_{j,t}, \tag{85}$$

where $y_{j,t}$ are the outcome variables of interest described below, $\mathbb{1}(E)_j$ is an indicator function equal to one if firm j in 2019 had no rental commitments in future years (and were therefore "exposed" to the remote work revolution) and zero otherwise. The term $\mathbb{1}(T)_{j,t}$ is an indicator equal to one in periods 2020-2023 and zero otherwise and the term $\mathbf{X}_{j,t}$ represents control variables which include sector, city and year fixed effects (see the Appendix for further details).[24]

As noted above, we will consider several outcome variables of interest. First and foremost, we consider rental expenditures per worker, $rent_{j,t}/n_{j,t}$, as a key measure of non-wage variable costs. In our model, this maps directly into $g(\omega)\kappa_n$. Second, following Gorodnichenko and Weber (2016), we use firms selling, general, and administrative expenses, $SGA_{j,t}$, as a measure of their fixed costs, $g(\omega)\kappa_o$. Third, while the wage bill is not separately reported in Compustat, it is a major part of the costs of goods sold, $COGS_{j,t}$. Therefore, we use information on these variable costs to proxy firms' wage bills, $h(\omega)Wn$. Finally, we consider firms labor productivity, $sales_{j,t}/n_{j,t}$, as a final variable of interest which in our model maps to $f(\omega)y/n$.

Note that while rental commitments directly affect firms' ability to change $rent/n$, they impact the remaining variables indirectly. This is because exposed firms are more likely to adopt remote work, that is, increase ω, because they can more easily reap its cost-saving benefits.

Impact of remote work on firm-level outcomes: Results. Table 8 provides the results. The coefficients, $\beta_{E,T}$, measure how costs and productivity of exposed firms changed between 2016-2019 and 2020-2023, i.e., during the onset of the remote work revolution, relative to changes observed in non-exposed firms. To the extent that our exposure measure captures differences in the uptake of remote work across firms, our theoretical results predict that $\beta_{E,T} < 0$.[25]

Indeed, the table shows that all coefficients are negative. Understandably, the coefficients on rental costs are strongest as our exposure measure is based on rental commitments. Nevertheless, the results suggests that exposed firms also experienced stronger declines in fixed and variable costs as well as a stronger drop in labor productivity compared to non-exposed firms. Moreover, with the exception of one specification all these effects are also statistically significant. These results, therefore, provide an independent

[24] Controlling for city and year fixed effects accounts for potential geographical heterogeneity and trends in rental prices, κ_n.

[25] In our model, the remote work revolution was associated with changes in structural parameters making work from home more efficient (an increase in $f(\omega)$). Despite this effect, the model would still predict $\beta_{E,T} < 0$ in the case of labor productivity, simply because of the strength of the rise in remote work rates.

Table 8: Estimation results

	rental costs		fixed costs		variable costs		productivity	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Exposure, $\beta_{E,T}$	-0.46^{***}	-0.55^{***}	-0.19^{**}	-0.10^{**}	-0.10	-0.09^{**}	-0.07^{*}	-0.06^{**}
	(0.05)	(0.07)	(0.08)	(0.04)	(0.08)	(0.04)	(0.04)	(0.03)
Fixed effects		✓		✓		✓		✓
R-squared	0.02	0.43	0.06	0.55	0.0002	0.50	0.01	0.41
Observations	23,006	23,006	23,018	23,018	25,666	25,666	25,139	25,139

Note: The table presents the coefficients of interest on exposed firms in periods 2020-2023 from regressions (85), $\beta_{E,T}$. Fixed effects include sector, city and year fixed effects. Standard errors (reported in brackets) are clustered at the firm level.

empirical validation of our model mechanisms.

The Appendix C.8 provides further details and robustness checks. These include specifications with firm fixed effects, an event study design as well as specifications considering the feasibility of remote work across industries based on Dingel and Neiman (2020). All these provide additional confirmation of our results. Moreover, we also consider a placebo test, estimating regressions (85) in the period 2010-2017 during which remote work did not experience such a strong increase. In this case, the coefficients $\beta_{E,T}$ are not statistically different from zero, validating our novel identification strategy.

6.3 Discussion

Our quantitative analysis is based on a new model in which heterogeneous businesses optimally choose the extent of remote work. In this subsection, we discuss some features of the current model and sketch potential extensions which may be fruitful avenues for future research.

Aggregate growth. In our framework, firms differ in the *level* of their long-run productivity. While these long-run differences are endogenous and indeed respond to changes in remote work conditions, for tractability we have abstracted from innovation and aggregate growth.

Some recent evidence (see Lin, Frey and Wu, 2023) suggests that remote collaboration may be linked to lower chances of breakthrough innovations. Therefore, in future research, it would be interesting to investigate how remote work interacts with innovation across heterogeneous firms – both at the intensive margin for a given set of incumbent businesses, and at the extensive margin, i.e., how remote work changes affect the incentives for the entry of innovative businesses. If the shift towards less productive firms identified in this paper would also lead to a decline in the innovative capacity of the economy, the welfare conclusions may be even less favorable.

Labor adjustment costs. While our model considers capital adjustment costs, future research may focus also on the costs of hiring and firing workers. More detailed data could inform researchers about how remote work affects the costs of attracting and retaining workers and how important these costs are for different types of firms.

On the one hand, the possibility of hiring workers remotely could loosen potential frictions in attracting (high-skilled) labor which may be locally scarce. On the other hand, while the costs of running a hiring process remotely may be lower, the efficiency of screening and information extraction could be reduced.

Agglomeration effects. While not considered in this paper, an additional channel through which remote work may impact *aggregate* outcomes is through agglomeration effects. For instance, using job posting wage data, Liu and Su (2024) show that remote work lead to a decrease in the urban productivity premium. They attribute this to "weakened agglomeration economies due to decreased interpersonal interactions." Within our framework, weakened agglomeration economies would further drag down aggregate productivity – over and above the impact of changes in business dynamism discussed in this paper.

Other factors and transition dynamics. Our model predicts that almost half of the entry rate spike during and in the aftermath of the COVID pandemic can be explained by increased remote work arrangements. Of course, other factors may have also contributed to the entry rate increase – e.g., the Payment Protection Program, or geographic restructuring of production in urban areas (see e.g., Decker and Haltiwanger, 2024).

Finally, let us note that the patterns of remote work are still evolving. While beyond the scope of this paper, it would be interesting to analyze the transition dynamics from the pre- to the post-pandemic worlds to gauge the timing of firm selection and the implications for aggregate outcomes in the medium run.

Cross-country comparison. A key result of our framework is that the strength of an increase in firm entry is indicative of the aggregate gains from the remote work revolution. In the Appendix, we show that while work from home rates spiked in a relatively similar fashion across many developed economies, changes in firm entry vary quite substantially. For instance, while Sweden has experienced a strong increase in firm entry – almost double that of the U.S. – Germany saw a *decline* in the number of startups. This suggests that institutional or other barriers to firm entry may play a key role in the ability of different countries to reap the benefits of remote work becoming cheaper, more efficient and more desirable.

7 Conclusion

In this paper, we study the macroeconomic impact of the large increase in work from home arrangements observed since the COVID-19 pandemic. We do so by proposing a new macroeconomic model of business dynamism in which firms can optimally choose to conduct part of their production remotely. We show analytically how such a framework generates a link between observed work from home rates, firm entry, and the composition of firms. Extending our baseline framework along several dimensions, we then quantify the macroeconomic impact of the remote work revolution and validate our model's mechanisms using a novel identification strategy.

We find that the observed rise in remote work rates can account for almost half of the firm entry rate increase since the COVID-19 pandemic. It also leads to an increase in output, consumption and welfare. However, these effects crucially depend on the strength and persistence with which firm entry increases. Indeed, if other frictions prevent business entry to rise sustainably, the shift towards smaller and less productive firms will dominate and eliminate all direct benefits of more favorable remote work.

Our paper also opens the door to several additional aspects which would be interesting to study. For example, how does remote work interact with other (e.g., financial or labor market) frictions? How may remote work arrangements affect firm-level and aggregate outcomes in the presence of two-sided heterogeneity and bargaining between workers and firms? We leave these and other questions for future research.

Chapter 3

Knowledge Generality, Competition and Growth

Chenchuan Shi
University of Oxford

Abstract

This paper studies how the *generality* of knowledge—its applicability across technologies and industries—shapes firms' innovation strategies, market structure, and aggregate growth. I build an endogenous growth model in which firms choose between general and firm-specific R&D while competing for market leadership. General innovations enhance firms' capacity to absorb and apply outside knowledge, creating spillovers within and across industries, whereas firm-specific innovations yield mainly private gains. The model predicts, and the data confirm, that (1) leaders favor firm-specific R&D while followers rely on general innovations to catch up, and (2) the gap in innovation generality between them follows a U-shaped pattern with market concentration. Leveraging variation in the enforceability of non-compete agreements across U.S. states, I provide empirical evidence consistent with the model's spillover mechanisms. The findings point to a novel growth policy: encouraging general R&D, particularly among leading firms, can improve knowledge diffusion and sustain long-run growth.

JEL codes: D22, L13, O30, O40.

University of Oxford, 10 Manor Rd, Oxford OX1 3UQ, UK, email: chenchuan.shi@nuffield.ox.ac.uk.
I am indebted to Matin Ellison and Petr Sedláček for the guidance and support throughout this project. I would like to thank Steve Bond, Timo Boppart, Andrea Chiavari, Paola Conconi, Sergio de Ferra, Maarten De Ridder, Tore Ellingsen, Xavier Jaravel, Alexandre Kohlhas, Hannes Malmberg, Michael McMahon, Silvia Miranda-Agrippino, Jane Olmstead-Rumsey, Morten O. Ravn, Abhijit Tagade, Alex Teytelboym, Florian Trouvain, Joshua Weiss, Zijun Xu, Francesco Zanetti and Wenchuan Zhao and conference and seminar participants of 17th Normac Symposium, Transatlantic Doctoral Conference, London Macro PhD Workshop, CEPR Long-Run Productivity Conference, MMF Annual Conference, Theories and Methods in Macro, Doctoral Workshop on Quantitative Dynamic Economics and University of Oxford for helpful discussions and comments.

1 Introduction

Innovation is a fundamental driver of long-run growth, but not all innovations are alike. Some are firm-specific, benefiting primarily the innovating firm. Others are general, creating spillovers across technologies and industries. This distinction—captured by the notion of *innovation generality* (Trajtenberg, Henderson and Jaffe, 1997), which measures the breadth of an idea's impact—is crucial for understanding how firms strategically respond to knowledge spillovers and, in turn, shape market competition and aggregate growth.

To study how firms choose between general and firm-specific innovations and the aggregate implications of their decisions, I develop a novel and tractable endogenous growth model. In this framework, two firms compete along a technological ladder. The focus on *innovation generality* introduces a key dimension for understanding firms' strategic responses to knowledge spillovers. General innovations, while benefiting rivals, also expand a firm's own capacity for learning and absorption by connecting it to a broader pool of ideas (Cohen and Levinthal, 1990).[1] An illustration comes from the photography industry: firms were able to invent and refine digital cameras by building on semiconductor technology, a general innovation developed outside their field (Fossum, 2020). The choice between general and firm-specific innovation is therefore inherently strategic: it balances the protection of a firm's current market position against the opportunity to exploit the expanding frontier of general knowledge.

More concretely, the model builds on an endogenous growth framework with step-by-step innovations in a leader–follower setting. It features two types of general knowledge spillovers: cross-industry and within-industry. By investing in general innovations, both leaders and followers gain access to cross-industry spillovers that enable them to absorb and apply technologies developed in other sectors. Crucially, these spillovers are not automatic; they require firms to engage in general R&D. Within each industry, spillovers flow from the leader to the follower, with their magnitude increasing in the leader's general R&D effort. Yet followers cannot benefit passively either; they also need to undertake general innovations to absorb and exploit the knowledge produced by leaders. This assumption about within-industry spillovers is later validated empirically through a quasi-natural experiment that leverages policy-driven variation across U.S. states with firm-level data.

The model incorporates a dynamic feedback loop between firms' innovation choices

[1] Importantly, though basic innovations also create broader spillovers, they differ from general innovations in several aspects: basic research is non-applied and mainly conducted by public institutions (e.g., universities) (Akcigit, Hanley and Serrano-Velarde, 2020), whereas general research is applied, largely conducted by private firms, and more prevalent in smaller and younger firms.

and the aggregate level of general knowledge. General R&D contributes to this shared knowledge base, which in turn expands learning opportunities and increases the returns to such investments. This encourages firms to pursue further general innovations. In this way, general innovations not only drive growth directly but also reinforce firms' incentives to invest in general R&D—a feedback mechanism largely overlooked in existing growth models.

Building on these mechanisms, the model yields two central theoretical predictions. First, leaders tilt toward firm-specific R&D while followers prefer general R&D: leaders seek to protect their dominance by limiting within-industry spillovers, while followers rely on general innovations to absorb external knowledge and accelerate catch-up. Second, the gap in innovation generality—defined as the share of R&D efforts devoted to general innovations—between the two firms varies non-monotonically with the level of market concentration. When firms compete neck-and-neck, they have the same level of innovation generality. Once a small leader–follower productivity gap emerges, followers have strong incentives to pursue general innovations, as the potential profit gain from overtaking leaders is large, while leaders focus on firm-specific innovations to defend against being displaced. However, when the gap is large, followers have weaker incentives to catch up, and leaders face less pressure to protect their position, narrowing the generality gap. This pattern implies that market structure does not only affect competition in the short run; it shapes the long-run direction of innovation, the extent of knowledge spillovers, and ultimately the pace of growth.

To test these predictions, I construct an empirical measure of firm-level innovation generality using the established indicator of patent generality (Trajtenberg et al., 1997). Patent generality captures the breadth of an innovation's impact: the extent to which it influences subsequent innovations across technological fields.[2] For example, a U.S. patent for a "location-aware application development framework" (granted in the early 2000s) has high generality (score of 0.83), as it enabled diverse applications across industries, from app-based taxi services and logistics to tourism, marketing, and healthcare. By contrast, a narrowly tailored invention such as the "stack clearing device and method" addresses a specific technical problem through a particular implementation and therefore has limited spillover potential and displays low generality (score of 0.18).

The measure of firm-level innovation generality is based on a dataset that integrates patent information with firm-level data from multiple sources. Patent-level data are sourced from the United States Patent and Trademark Office (USPTO) and PatentsView, and are linked to Compustat, with standard firm-level information. The mapping between

[2]Formally, patent generality is defined as $1-\sum_j s_{ij}^2$, where s_{ij} denotes the share of citations to patent i originating from technological class j (Trajtenberg et al., 1997).

patents and firm assignees follows Kogan, Papanikolaou, Seru and Stoffman (2017), which provides carefully constructed firm–patent matches. Using average patent generality at the firm level—referred to as innovation generality—as the empirical counterpart to the model, the analysis confirms its key theoretical predictions: Leaders produce less general innovations than followers, and the gap in innovation generality between them follows a U-shaped relationship with market concentration.

To further validate the model's key assumption of within-industry spillovers, I use a quasi-natural experiment based on variation in state-level enforceability of non-compete agreements. These agreements affect the mobility of R&D workers, which is an important channel of knowledge diffusion (Jaffe, Trajtenberg and Henderson, 1993; Almeida and Kogut, 1999; Singh and Agrawal, 2011; Stoyanov and Zubanov, 2012; Liu, 2023). Stronger enforcement restricts inventor mobility and thereby reduces the efficiency of within-industry knowledge spillovers. Building on this mechanism, the empirical results show that stricter enforcement encourages industry leaders to produce more general innovations, while discouraging followers from doing so. This aligns with the model's logic: when within-industry spillovers are weakened, leaders shift from firm-specific to general innovations because they face less risk of being overtaken, while followers lose incentives to pursue general innovations as a way to absorb external general knowledge and catch up.

To quantify the effects of innovation generality on growth and study its policy implications, I calibrate the model using U.S. data from 1995–2000, a period characterized by high innovation generality and peak economic growth over the past four decades. The comparative statics yield two key insights. First, reducing the cost of general R&D stimulates growth more effectively than reducing firm-specific R&D cost, as it magnifies spillovers both within and across industries. Second, when within-industry spillovers become highly efficient, aggregate growth may decline, as leaders are strongly discouraged from investing in general innovations. This reduction offsets followers' expansion of general R&D and limits the spillovers available to firms in other industries. Consequently, weak enforcement of non-compete agreements may not necessarily promote growth: if inventors carrying general knowledge can easily move to follower firms, leaders may have little incentive to undertake general innovations in the first place.

Building on the comparative statics, I evaluate the effects of general R&D subsidies. The key finding is that subsidizing leaders' general innovations is typically more effective in promoting aggregate growth than supporting followers', especially in industries where general R&D efficiency is high.[3] Targeting leaders' general R&D strengthens their incentives to pursue general knowledge, which in turn encourages followers to expand

[3]That is, where the cost scale of general R&D is low.

their own R&D to absorb it. This increases the aggregate level of general knowledge and further enhances cross-industry spillovers through a feedback loop. By contrast, when subsidies are directed to followers, they increase their general R&D to catch up, but leaders in response scale back their own efforts to avoid being displaced, partly offsetting the aggregate gains.

This policy implication differs from the traditional leader–follower framework in the literature (see, e.g., Liu, Mian and Sufi, 2022; Akcigit and Ates, 2023), where supporting followers is typically viewed as the optimal policy. While subsidizing followers remains valuable—since they are often self-motivated to pursue general innovations and require less monitoring—policies that sharpen leaders' incentives to invest in general R&D can create larger spillovers and stimulate growth more effectively. Practical tools therefore include prize systems that explicitly reward leading firms for developing general innovations and research collaborations that encourage cross-sector knowledge sharing.[4]

Finally, the model provides a lens on the link between innovation generality and secular trends. Since 2000, the U.S. has experienced a marked growth slowdown, coinciding with a sharp decline in aggregate patent generality across most major sectors. Calibrating the model to 2010–2015 data suggests that reduced efficiency of general R&D investment explains about half of the observed growth slowdown, as well as a large fraction of the decline in firm entry and leadership turnover. As general ideas become harder to find, firms shift their focus toward firm-specific innovations, which hinders knowledge diffusion and makes it more difficult for followers to catch up. In other words, slower technological diffusion and lower leadership turnover emerge endogenously from the lower efficiency of general R&D faced by both leaders and followers. In this way, the model connects falling innovation generality to slower diffusion and lower business dynamism—highlighting how long-run growth and market structure are jointly shaped by the scope of innovation.

Related Literature. This paper relates to several strands of literature. First, it contributes to work on heterogeneous innovation and firm dynamics (Klette and Kortum, 2004; Acemoglu and Cao, 2015; Akcigit and Kerr, 2018; Garcia-Macia, Hsieh and Klenow, 2019; Caggese, 2019; Peters, 2020; Acemoglu, Akcigit and Celik, 2022; Caicedo and Pearce, 2024; Casal, 2025; Fernández-Villaverde, Yu and Zanetti, 2025). Prior studies distinguish between radical vs. incremental, external vs. internal, and basic vs. applied innovations, as well as differences between incumbents and entrants. Building on this tradition, I focus on innovation generality, which captures the extent of knowledge spillovers across firms. Importantly, while both general and basic innovations create

[4]Recent work shows that cash prizes can complement subsidies by steering innovation toward socially valuable outcomes (see, e.g., Che, Iossa and Rey, 2021; Graff Zivin and Lyons, 2021).

broad spillovers, they differ in several aspects: basic research is non-applied and mainly conducted by public institutions (e.g., universities) (Akcigit et al., 2020), whereas general research is applied, largely conducted by private firms, and more prevalent in smaller and younger firms.

Second, this paper relates to the literature on technological spillovers (Jaffe et al., 1993; Jaffe, Trajtenberg and Fogarty, 2000; Bloom, Schankerman and Van Reenen, 2013; Perla and Tonetti, 2014; Arora, Belenzon and Sheer, 2021; Benhabib, Perla and Tonetti, 2021; Liu and Ma, 2021; Dyévre, 2024; Giroud, Liu and Mueller, 2024). Earlier work emphasizes agglomeration effects, where inventor productivity rises with the number of inventors in the same field or location. General innovations, however, create broader spillovers across technological boundaries. To capture this, I introduce spillover mechanisms both within and across industries. Unlike research on patenting strategies to block rivals (Argente, Baslandze, Hanley and Moreira, 2023; Cunningham, Ederer and Ma, 2021), my focus is on knowledge spillovers—captured through innovation generality—as a strategic margin.

Finally, the paper contributes to the literature on the recent U.S. productivity slowdown (Liu et al., 2022; Aghion, Bergeaud, Boppart, Klenow and Li, 2023; De Ridder, 2024; Olmstead-Rumsey, 2025). Existing research highlights weaker spillovers from leaders or declining innovation quality of followers as mechanisms that widen firm inequality and dampen growth (Akcigit and Ates, 2023; Olmstead-Rumsey, 2025). Different from prior studies, this paper focuses on the role of general R&D efficiency in shaping firms' strategic innovation decisions and in driving the slowdown in growth and business dynamism.

The remainder of the paper is organized as follows. Section 2 develops the theoretical framework and derives predictions about knowledge generality. Section 3 introduces the empirical measure of innovation generality. Section 4 quantifies the model and tests its predictions. Section 5 examines the policy implications and the role of knowledge generality in driving the post-2010 growth slowdown. Section 6 concludes.

2 A Growth Model with Knowledge Generality

To study how firms adjust their innovation strategies in response to knowledge spillovers and the aggregate implications of their decisions, I develop an endogenous growth model in which firms choose between general and firm-specific innovations. The model is set in continuous time and features a continuum of heterogeneous markets, which differ in the efficiency of general R&D investment. In each market, two firms compete along a technological ladder, following the framework of Aghion, Harris, Howitt and Vickers

(2001), Liu et al. (2022), and Akcigit and Ates (2023). I begin by describing household preferences, then outline firms' pricing and innovation decisions, define the equilibrium, and finally present the theoretical predictions about innovation generality.

2.1 Household Preferences

The representative household owns firms in the economy and provides production workers L inelastically. At each instance t, the representative household decides its consumption across a continuum of heterogeneous duopoly markets indexed by j, and maximizes their utility:

$$\max_{y_a(t,j), y_b(t,j)} \exp\Big\{ \int_0^1 \ln[y_a(t,j)^{\frac{\sigma-1}{\sigma}} + y_b(t,j)^{\frac{\sigma-1}{\sigma}}]^{\frac{\sigma}{\sigma-1}} dj \Big\}$$

$$\text{s.t.} \int_0^1 \big[p_a(t,j) y_a(t,j) + p_b(t,j) y_b(t,j)\big] dj = \Pi(t) + W(t)L \tag{86}$$

where $y_i(t,j)$ is the demand for firm i's product in market j, $p_i(t,j)$ is the price of firm i's product, with $i \in \{a, b\}$. The term $W(t)$ is the equilibrium wage of production workers, and $\Pi(t)$ represents the aggregate economy-wide profits.[5] The elasticity of substitution across the two varieties within each market is measured by σ.

2.2 Firms

In each market, two incumbent firms compete for leadership. I first describe their pricing decisions and then their innovation choices.

Pricing. Firm i in market j has the following production function:

$$y_i(t,j) = \gamma^{z_i(t,j)} l_i(t,j) \tag{87}$$

where $y_i(t,j)$ is firm i's output, $\gamma^{z_i(t,j)}$ is firm i's productivity, and $l_i(t,j)$ is the number of production workers it employs. Firms improve their productivity through step-by-step innovations, with γ the innovation step size. Henceforth, the indices j and t are dropped to simplify the notation. In each market, the two firms engage in Bertrand competition

[5]It is important to note that the R&D investments are paid directly from firm profits, $\Pi(t)$, and are paid back to the representative households. Hence they do not appear on the right hand side of budget constraint (86).

by setting their prices p_i and solve:

$$\max_{p_i}(p_i - W\gamma^{-z_i})y_i$$
$$\text{s.t. } p_a y_a + p_b y_b = \Pi + WL \ \text{ and } \ y_a/y_b = (p_a/p_b)^{-\sigma} \tag{88}$$

The pricing problem can be analyzed based on the gap between the leader and the follower, denoted as $s = |z_a - z_b|$. When $s > 0$, one firm acts as a temporary leader while the other follower. If $s = 0$, the two firms are considered to be neck-and-neck. Solving (88) yields the price ratio $\rho_s = p_s/p_{-s}$ between the leader and the follower, which satisfies:

$$\rho_s^\sigma = \gamma^{-s}\frac{\sigma\rho_s^{\sigma-1}+1}{\sigma+\rho_s^{\sigma-1}} \tag{89}$$

Let π_s represent the leader's profit, normalized by the total income $(\Pi + WL)$, and π_{-s} the normalized follower's profit. We then have a pair of profits fully characterized by the distance s:

$$\pi_s = \frac{\rho_s^{1-\sigma}}{\sigma+\rho_s^{1-\sigma}}, \ \pi_{-s} = \frac{1}{\sigma\rho_s^{1-\sigma}+1} \tag{90}$$

This pair of profits indicates that the leader earns more profit than the follower for all $s \geq 1$, and its profit increases with s.

Innovation Types: General and Firm-Specific. To improve their productivity, firms allocate their resources between two types of innovations, general and firm-specific. Unlike other forms of heterogeneous innovations[6], general and firm-specific innovations differ in their contribution to knowledge spillovers and how they shape a firm's capacity to absorb outside ideas.

General innovations create spillovers that are widely accessible, benefiting not only the innovating firm but also its competitors and even firms in other industries. Firms nevertheless still have strong incentives to pursue them, as general innovations expand absorptive capacity—the ability to recognize, learn from and build on external general knowledge. By contrast, firm-specific innovations are narrowly tailored to a firm's own operations, products, or processes. Their benefits remain largely internal and are difficult for competitors to imitate or appropriate.

These contrasting features—broad spillovers and absorptive gains from general innovation versus narrow, private returns from firm-specific innovation—play a central role

[6]For example, radical versus incremental, external versus internal, and basic versus applied (Klette and Kortum, 2004; Acemoglu and Cao, 2015; Akcigit and Kerr, 2018; Garcia-Macia et al., 2019; Caggese, 2019; Peters, 2020; Acemoglu et al., 2022).

in shaping firms' strategic choices in response to knowledge spillovers.

Innovation Decisions. I use the superscript (or subscript) k to index different types of markets, which differ in the efficiency of general R&D investment.[7]

A firm that is currently in the leadership position incurs a firm-specific R&D cost $c_\eta(\eta_s^k)$ and a general R&D cost $c_\lambda^k(\lambda_s^k)$ in exchange for a Poisson rate $(\eta_s^k + \Phi\lambda_s^k)$ to improve its productivity z_s by one step. Here, η_s^k and λ_s^k represent the leader's firm-specific and general R&D efforts, respectively. Similarly, the follower chooses its own firm-specific and general R&D efforts, denoted by η_{-s}^k and λ_{-s}^k. Both types of effort contribute to firm-level productivity, but their growth payoffs differ.

I introduce two forms of spillovers arising from general innovations: cross-industry and within-industry spillovers.

Cross-industry Spillovers. By investing in general innovations, both firms gain access to cross-industry knowledge spillovers, captured by Φ, which is determined endogenously by the aggregate level of general R&D efforts in equilibrium:

$$\Phi \equiv \Phi(\Lambda) \tag{91}$$

where Λ is the aggregate level of general R&D efforts of all incumbents across all types of markets.

Importantly, these spillovers are not freely available: firms need to undertake general innovations themselves in order to absorb and apply the technologies developed in other industries.[8]

Within-industry Spillovers. Within an industry, spillovers flow from the leader to the follower, and their magnitude increases with the leader's general R&D effort, captured by $f(\lambda_s^k)$. However, as with cross-industry spillovers, followers cannot benefit passively: they should also invest in general innovations to absorb and exploit the knowledge created by the leader.

[7] Specifically, higher efficiency of general R&D investment corresponds to a lower cost scale.

[8] As an example of the cross-industry spillovers, artificial intelligence (AI) tools were initially invented and advanced mainly within high tech sectors. Then these tools (e.g., natural language processing and machine learning) were adopted and integrated by other sectors, such as finance (e.g., fraud detection), healthcare (e.g., medical imaging), manufacturing (e.g., automation), education (e.g., adaptive learning platforms), entertainment (e.g., video recommendations), etc.

The state $s(k)$ transitions over time interval Δ according to:

$$s(k, t+\Delta) = \begin{cases} s(k,t)+1 & \text{with probability } \Delta \cdot (\eta_s^k + \Phi\lambda_s^k), \\ s(k,t)-1 & \text{with probability } \Delta \cdot (\eta_{-s}^k + \Phi f(\lambda_s^k)\lambda_{-s}^k + h), \\ s(k,t) & \text{otherwise.} \end{cases}$$

In the first line, the term $\Phi\lambda_s^k$ indicates that the leader's general knowledge production benefits from cross-industry spillovers, Φ. In the second line, the term h is the fixed catch-up rate for the follower, a standard and important assumption to ensure the convergence of the model. The term $\Phi f(\lambda_s^k)\lambda_{-s}^k$ reflects that the follower's general innovations benefit from two sources: cross-industry spillovers, Φ, and within-industry spillovers, $f(\lambda_s^k)$, from the leader. As the leader invests more in general knowledge production, the follower gains greater resources to learn and imitate. However, to effectively absorb this external general knowledge, the follower should also engage in general innovations. In Section 4, I empirically validate the spillover mechanism through a quasi-natural experiment that leverages policy-driven variation across U.S. states together with firm-level data.

Denote v_s^k, v_{-s}^k and v_0^k as the normalized value functions for the leader, the follower and the neck-and-neck firms.[9] Then the dynamic optimization problem is independent of time t, and can be characterized by the following HJB equations:

$$\begin{aligned} rv_s^k &= \max_{\eta_s^k, \lambda_s^k \geq 0} \{\pi_s - c_\eta(\eta_s^k) - c_\lambda^k(\lambda_s^k) + [\eta_{-s}^k + \Phi f(\lambda_s^k)\lambda_{-s}^k + h](v_{s-1}^k - v_s^k) \\ &\quad + (\eta_s^k + \Phi\lambda_s^k)(v_{s+1}^k - v_s^k) - \tau^k v_s^k\} \quad (92) \\ rv_{-s}^k &= \max_{\eta_{-s}^k, \lambda_{-s}^k \geq 0} \{\pi_{-s} - c_\eta(\eta_{-s}^k) - c_\lambda^k(\lambda_{-s}^k) + [\eta_{-s}^k + \Phi f(\lambda_s^k)\lambda_{-s}^k + h](v_{-s+1}^k - v_{-s}^k) \\ &\quad + (\eta_s^k + \Phi\lambda_s^k)(v_{-s-1}^k - v_{-s}^k) - \tau^k v_{-s}^k\} \quad (93) \\ rv_0^k &= \max_{\eta_0^k, \lambda_0^k \geq 0} \{\pi_0 - c_\eta(\eta_0^k) - c_\lambda^k(\lambda_0^k) + (\eta_{-0}^k + \Phi\lambda_{-0}^k)(v_{-1}^k - v_0^k) \\ &\quad + (\eta_0^k + \Phi\lambda_0^k)(v_1^k - v_0^k) - \tau^k v_0^k\} \quad (94) \end{aligned}$$

where r is the growth-adjusted interest rate. The competitor's R&D decisions in state 0 (neck-and-neck competition) are given by η_{-0}^k for firm-specific R&D and λ_{-0}^k for general R&D. Both firms face an endogenous creative destruction rate τ^k, which is determined in equilibrium and depends on market type.

[9] Consistent with the firms' profits (π_s and π_{-s}), the value functions (v_s^k, v_{-s}^k and v_0^k) are normalized by the total income ($\Pi + WL$). Note that along the balanced growth path, they grow at the same rate.

2.3 Entrants

There is a continuum of potential entrants who conduct general R&D and take over a random incumbent in each market. Upon entry, an entrant inherits the productivity level (z) of the incumbent it replaces. Because entrants' innovations are assumed to be inherently general, they can build on the aggregate stock of general knowledge produced by incumbents.[10] Thus, entrants benefit from cross-industry spillovers in the same way as incumbents do through their general innovations. A higher degree of cross-industry spillovers (Φ) therefore leads to a higher entry rate. To achieve an entry rate of $\Phi\lambda_e$, entrants incur an R&D cost:

$$c_e(\lambda_e) = \frac{1}{\theta}(\kappa_e\lambda_e)^\theta \tag{95}$$

Upon entry, entrants have a probability p_k of entering a market of type k. This probability is known and endogenously determined in equilibrium. The optimization problem for entrants becomes:

$$\max_{\lambda_e} \ -c_e(\lambda_e) + \Phi\lambda_e \sum_k p_k \mathbb{E}\big[\frac{1}{2}(v_s^k + v_{-s}^k)\big] \tag{96}$$

Hence, the optimal entry rate λ_e satisfies the following condition:

$$\Phi \sum_k p_k \mathbb{E}\big[\frac{1}{2}(v_s^k + v_{-s}^k)\big] = \kappa_e^\theta \lambda_e^{\theta-1} \tag{97}$$

2.4 Equilibrium Definition

The equilibrium is characterized by market-dependent creative destruction rates $\{\tau^k\}$, entrants' R&D effort λ_e, the endogenous cross-industry spillover effect Φ, an infinite collection of value functions, and R&D efforts $\{v_s^k, v_{-s}^k, \eta_s^k, \eta_{-s}^k, \lambda_s^k, \lambda_s^k\}_{s=0}^\infty$ such that: The incumbents' pricing decisions satisfy (88) and generate $\{\pi_s, \pi_{-s}\}_{s=0}^\infty$; The value functions solve the HJB equations (92) through (94); Entrants' decisions satisfy (97); The creative destruction rate satisfies: $\tau^k = \frac{p_k\Phi\lambda_e}{2\chi_k}$, where χ_k is the mass of market of type k; The cross-industry spillover is determined in equilibrium according to equation (91), by the aggregate level of general knowledge produced by incumbents, which can be expressed as:

$$\Lambda = \frac{1}{2}\sum_k \sum_{s=0}^\infty \chi_k \mu_s^k (\lambda_s^k + \lambda_{-s}^k) \tag{98}$$

Denote the steady-state probability distribution of the productivity gaps in the type-k

[10] In the Appendix, I show that empirically entrants exhibit higher innovation generality than incumbents.

market as $\{\mu_s^k\}_0^\infty$. The aggregate growth rate g satisfies that for all k:

$$
\begin{aligned}
g &= \ln(\gamma)\Big[\sum_{s=1}^{\infty}\mu_s^k(\eta_s^k+\Phi\lambda_s^k)+2\mu_0^k(\eta_0^k+\Phi\lambda_0^k)\Big] \\
&= \ln(\gamma)\Big\{\sum_{s=1}^{\infty}\mu_s^k\big[\eta_{-s}^k+\Phi f(\lambda_s^k)\lambda_{-s}^k+h\big]\Big\} \qquad (99)
\end{aligned}
$$

where the transition of the states satisfies:

$$
\begin{aligned}
2\mu_0^k(\eta_0^k+\Phi\lambda_0^k) &= \big[\eta_{-1}^k+\Phi f(\lambda_1^k)\lambda_{-1}^k+h\big]\mu_1^k \\
\mu_s^k(\eta_s^k+\Phi\lambda_s^k) &= \big[\eta_{-(s+1)}^k+\Phi f(\lambda_{s+1}^k)\lambda_{-(s+1)}^k+h\big]\mu_{s+1}^k
\end{aligned}
$$

Equation (99) states that in the steady state, the average growth rate of the leader and neck-and-neck firms should equal the average growth rate of the follower in each type of the market. Along the balanced growth path, all types of markets have the same average growth rate.

2.5 Model Predictions

The model yields two predictions linking leadership, innovation generality, and market concentration, which I test empirically in Section 4. From a theoretical perspective, these predictions also distinguish general and firm-specific innovations from other dimensions of heterogeneous innovation.

Definition: Innovation Generality. I define firm-level innovation generality as the share of general R&D effort in total R&D efforts. For leaders, this is given by:

$$
\psi_s^k = \frac{\lambda_s^k}{\lambda_s^k+\eta_s^k} \qquad (100)
$$

This measure captures, on average, the breadth of impact of the ideas produced by a firm. A higher value of ψ_s^k (ψ_{-s}^k for the follower) indicates that a larger fraction of the leader's innovation effort is directed toward general knowledge, which is then more widely applicable across firms and industries. This concept can be linked to the empirical measure documented in Section 3.

Proposition **7**

Suppose the cost functions are given by:

$$
c_\eta(\eta)=\tfrac{1}{\theta}(\alpha\eta)^\theta \quad \textit{and} \quad c_\lambda^k(\lambda)=\tfrac{1}{\theta}(\kappa_k\lambda)^\theta,
$$

with $\theta > 1$.[11] *Assume further that the within-industry spillover function* $f(\cdot)$ *satisfies* $f(x) \geq 1$ *and* $f'(x) > 0$ *for all* $x \geq 0$. *Then the innovation generality of the follower is higher than that of the leader, i.e.,*

$$\psi_{-s}^k > \psi_s^k \qquad \forall\, s \geq 1.$$

Proof. See Appendix A.1. □

Proposition 7 states that the follower prioritizes general innovations, while the leader prefers firm-specific innovations. This is because engaging in general innovations enhances the follower's capacity of learning and absorbing the general knowledge created by the leader, thereby increasing its probability of catching up. Conversely, while producing general knowledge contributes to the leader's growth and its absorptive capacity for cross-industry spillovers, it also raises the risk of being overtaken by the follower. Therefore, the leader has weaker incentives to invest in general innovations.

This prediction captures the differences in firms' investment patterns between general and basic innovations. Although both types generate greater knowledge spillovers, larger firms tend to invest less in general innovations, while only the very largest firms engage in basic research (Akcigit et al., 2020).

Corollary **5**

In the case of neck-and-neck competition (i.e., $s = 0$*), the innovation generality of both firms is given by:*

$$\psi_0^k = \frac{(\Phi\alpha^\theta)^{\frac{1}{\theta-1}}}{(\Phi\alpha^\theta)^{\frac{1}{\theta-1}} + \kappa_k^{\frac{\theta}{\theta-1}}}$$

As the technology gap between the leader and the follower diverges, the innovation generality of both firms converges to the same value:

$$\psi_s^k, \psi_{-s}^k \to \frac{(\Phi\alpha^\theta)^{\frac{1}{\theta-1}}}{(\Phi\alpha^\theta)^{\frac{1}{\theta-1}} + \kappa_k^{\frac{\theta}{\theta-1}}} \quad \textit{when } s \to \infty$$

Proof. See Appendix A.2. □

Corollary 5 states that while the follower's innovation generality is higher than that of the leader (as shown in Proposition 7), the gap between them diminishes as the distance between the leader and the follower diverges, and each firm's innovation generality approaches the level in the neck-and-neck competition. Intuitively, when the distance is

[11] Convex cost functions are common in the literature (see, e.g., Bloom, Griffith and Reenen (2002); Acemoglu, Akcigit, Alp, Bloom and Kerr (2018)).

small but positive, the follower has strong incentives to produce general knowledge, as the potential profit gain is great if the follower successfully catches up with the leader. In contrast, the leader has strong incentives to focus on firm-specific knowledge, as the potential profit loss from being overtaken is significant. This strategic interaction between the two firms towards general knowledge spillovers implies the following Proposition, which will be tested empirically in Section 4.

PROPOSITION 8
Define the innovation generality gap as $g_s := \psi_s^k - \psi_{-s}^k$, $s \in \mathbb{N}_0$. *Then the innovation generality gap between the two firms varies non-monotonically with market concentration. Formally, there exist* $a, b, c \in \mathbb{N}_0$ *with* $a < b < c$ *such that:*

$$g_a > g_b \quad \textit{and} \quad g_b < g_c$$

Proof. See Appendix A.3. □

This prediction captures the differences between general and external innovations along the dimensions of firms' R&D allocation and market concentration. The gap in innovation generality varies non-monotonically with market concentration, while the gap in the share of external product lines relative to total product lines decreases monotonically, as larger firms invest disproportionately less in external innovations (Akcigit and Kerr, 2018).

3 Knowledge Generality in the Data

In this section, I introduce an empirical measure of innovation generality, defined as the average breadth of impact of the ideas produced by a firm. To implement this concept, I draw on the established indicator of patent generality (Trajtenberg et al., 1997). While the notion of patent generality was first formalized in 1997, its implications for market competition and economic growth remain relatively underexplored. I begin by outlining the data sources used to measure innovation generality, and then extend its definitions to the firm level.

3.1 Data Sources

I compile a firm-level dataset that integrates information on patents and innovating firms from multiple sources. The core patent-level data come from the United States Patent and Trademark Office (USPTO) and PatentsView.[12] These patent data are linked to Compustat, which contains comprehensive firm-level information. The linkage between patent and firm assignees is based on Kogan et al. (2017), which provides carefully constructed firm–patent matches.[13]

Patent Data. Patent data were sourced from the USPTO and PatentsView. The USPTO provides information on patent application and grant years. Details on patent assignees, citations, and their subclasses were obtained from PatentsView. The analysis focuses on utility patents, which capture the technical innovation and functional aspects of inventions, in contrast to design patents that protect aesthetic designs, and plant patents that protect new plant varieties.

Innovating Firms. To construct a firm sample, I merge the patent dataset with Compustat through the data repository from Kogan et al. (2017), which provides a standardized linkage between firm identifiers across datasets. Firm-level financial and operational variables are obtained from Compustat, while patent-level data allow the calculation of innovation measures, such as generality, at the firm level. The sample covers the period from 1980 to 2018 and includes firms that filed at least one patent in a given year. The sample further excludes agricultural (SIC code 0100-0999), utility (SIC code 4900-4949), financial (SIC code 6000-6799), and non-operating establishments (SIC code 9995).

3.2 Innovation Generality: Construction and Definitions

I use patent generality to construct an empirical firm-level proxy for the model's concept of innovation generality. I begin by defining patent generality and illustrating its construction with a concrete example. This measure is then extended to represent innovation generality at the firm level.

Definition: Patent Generality. Patent generality measures the breadth of a patent's impact. A higher generality score indicates that the patent influenced subsequent innovations in a wider variety of fields, while a low generality score means that most citations are concentrated in a few fields.

Formally, patent generality is defined as

[12]PatentsView is an open platform which provides disambiguated data linking patents to unique inventor and assignee entities, and can be accessed through https://patentsview.org.

[13]The data from Kogan et al. (2017) can be accessed at https://github.com/KPSS2017/Technological-Innovation-Resource-Allocation-and-Growth-Extended-Data.

$$\psi_i = 1 - \sum_j s_{ij}^2 \tag{101}$$

where s_{ij} is the share of citations to patent i coming from patents in technology class j (Trajtenberg et al., 1997). To ensure time consistency, the generality measure is based on citation information within a 5-year window following a patent's grant.[14] In calculating this measure, self-citations are excluded in order to focus on the knowledge spillovers between firms.

Example: Generality Measure Construction. As an example of patent generality construction, the patent with ID 6594666 was granted in 2003 and assigned to Oracle Corporation (CRSP permno 10104) with SIC code 7370 (service-computer programming, data processing). The patent was about "location-aware application development framework", with detailed information shown in Figure 17. Within a 5-year window since the patent grant, it received citations from 63 utility patents across 10 technological classes defined according to Cooperative Patent Classification (CPC) subclasses, as shown in Table 9.[15] The generality score of patent 6594666 is then 0.829, suggesting a relatively broad scope of impact and, therefore, a more general innovation.

By comparison, the patent with ID 6550058, granted in the same year and assigned to International Business Machines Corporation (CRSP permno 12490) in the same industry as Oracle Corporation, was about "stack clearing device and method". It received citations from 10 utility patents spanning only 2 CPC subclasses, resulting in a generality score of 0.180.[16] This indicates that the patent's impact is relatively narrow in scope and therefore more firm-specific.

Definition: Firm-level Innovation Generality. Patent generality is then aggregated to the firm-year level as an indicator of firm innovation generality, by averaging among all patents that are produced by firm f in year t:

$$\psi_{f,t} = \frac{1}{n} \sum_{i=1}^{n} \psi_{i,t} \tag{102}$$

where $\psi_{i,t}$ is the generality score of patent i which was applied for in year t and assigned

[14] Note that the definition of generality requires a patent to have at least one forward citation within five years of its grant. In the Appendix, I show that the empirical results in Section 4.2 remain robust when patents with no forward citations are instead assigned a generality score of 0.

[15] In the patent sample, there are 672 CPC subclasses in total.

[16] The detailed patent information and generality score construction can be found in the Appendix.

Figure 17: Detailed Patent Information: Patent ID 6594666

(12) **United States Patent**
Biswas et al.

(10) **Patent No.: US 6,594,666 B1**
(45) **Date of Patent: Jul. 15, 2003**

(54) **LOCATION AWARE APPLICATION DEVELOPMENT FRAMEWORK**

(75) Inventors: **Prabuddha Biswas**, Nashua, NH (US); **Raja Chatterjee**, Nashua, NH (US)

(73) Assignee: **Oracle International Corp.**, Redwood Shores, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 234 days.

(21) Appl. No.: **09/669,503**

(22) Filed: **Sep. 25, 2000**

(51) **Int. Cl.**[7] **G06F 7/00**; G06F 17/00
(52) **U.S. Cl.** **707/100**; 707/102
(58) **Field of Search** 707/100, 102

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,948,040 A * 9/1999 DeLorme et al. 701/201

OTHER PUBLICATIONS

Oracle Spatial User's Guide and Reference, Release 8.1.6, Feb. 1999, 322 pages, Oracle Corporation (Oracle Part No. A77132–01) USA.
Oracle8i interMedia Audio, Image, and Video User's Guide and Reference, Release 8.1.5, Feb. 1999, 616 pages, Oracle Corporation (Oracle Part No. A67299–01) USA.
Oracle8i intermedia Locator, User's Guide and Reference, Release 8.1.5, Feb. 1999,, 42 pages, Oracle Corporation (Oracle Part No. A67298–01) USA.

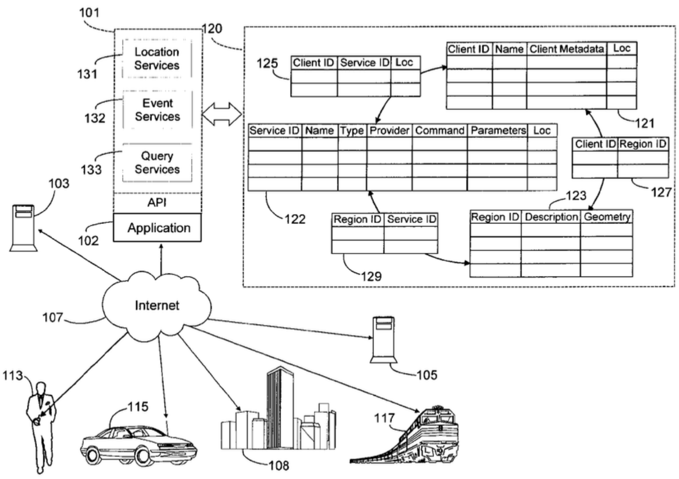
* cited by examiner

Primary Examiner—Safet Metjahic
Assistant Examiner—Cindy Nguyen
(74) *Attorney, Agent, or Firm*—Sanjay Prasad

(57) **ABSTRACT**

A shareable application program interface (API) infrastructure which is used in combination with a relational database to provide data storage and processing functions for location-aware objects, including particularly mobile objects whose current position is periodically updated by a position determining system. Client and service tables in the relational database are used to store the current point location, and other data, representing virtual objects, including mobile objects. A region table stores that describing the geometry and characteristics of geographical regions having defined boundaries within which the client and service objects reside. For each client, the set of services used by that client is recorded in a client profile database table. The services available on the system which are position-dependent have a geographical location associated with them. The API makes available an assortment of location dependent processing functions which may be used by location aware applications.

22 Claims, 2 Drawing Sheets



to firm f.[17] This measure of firm-level innovation generality serves as an empirical proxy for its theoretical counterpart in the model, that is, ψ_s^k for the leader and ψ_{-s}^k for the follower.

Note that the measure of firm-level innovation generality in (102) is the simple average of patent generality scores. In the Appendix, I construct a cost-adjusted version of this measure, using the real value of innovation from Kogan et al. (2017) to approximate patent cost (e.g., R&D expenditure or effort). The empirical results in Section 4.2 remain robust under this alternative specification.

[17] I use the application year rather than the grant year to more precisely capture the timing of knowledge production.

Table 9: Generality Score Construction: Patent 6594666

CPC subclasses	Definitions	Citations
$G01C$	measuring, surveying, navigation, gyroscopic instruments	8
$G06F$	electric digital data processing	19
$G06Q$	ICT for admin., comm., fin., manag. or superv. purposes	3
$G08G$	traffic control system	3
$G09B$	signaling or calling systems	1
$H04B$	transmission	1
$H04L$	transmission of digital information	6
$H04M$	telephonic communication	4
$H04W$	wireless communication networks	12
$Y10S$	technical subjects covered by former XRACs and digests	6

Note: The first and second column list the CPC subclasses the citing patents are from and their definitions. The third column lists the number of citing patents that belong to the subclass.

4 Parameterization and Model Validation

In this section, I parameterize the model and test its predictions on the relationship between leadership, innovation generality, and market concentration, using the empirical measure of knowledge generality developed in the previous section. Then I provide evidence supporting the mechanism of within-industry general knowledge spillovers, thereby validating a central assumption of the model.

4.1 Parameterization

The baseline model is calibrated to the 1995-2000 period, a time characterized by high levels of innovation generality and total factor productivity growth in the US. In Section 5, I compare the baseline model with the counterfactual, where general ideas become harder to find to reflect the post-2010 economy, marked by a lower aggregate generality and lower growth. Table 10 provides an overview of the parameter values and their sources. The first part of the table lists parameters set externally, and the second part shows parameters set via internal calibration. For tractability, I assume that there are two types of industries that differ in their efficiency in conducting general R&D investments, corresponding to the type-k markets in the model. Specifically, industries whose average patent generality between 1995 and 2000 lies above the median across all industries are classified as having high general R&D efficiency, while those with an average patent generality below the median are classified as having low efficiency. The superscript H represents high-type (in general R&D efficiency), while L denotes low-type.

Data Sources. The calibration relies on data from several key sources. Data on the aggregate growth rate are calculated based on Fernald (2015). Data on firm entry are from the Business Dynamics Statistics (BDS). Firm-level innovation generality and R&D to sales ratio, as well as the turnover rate of industry leadership, are calculated by linking patent data from the USPTO and PatentsView, with Compustat firm data, based on the KPSS data repository. The National Center for Science and Engineering Statistics (NCSES) provides information on the R&D to GDP ratio. To ensure consistency, all moments used in the calibration are calculated as averages over the period of 1995–2000.

Functional Forms. Two functions govern the efficiency of cross- and within-industry knowledge spillovers, $\Phi(\cdot)$ and $f(\cdot)$. Following León-Ledesma and Satchi (2019) in their analysis of technology adjustment and balanced growth, I adopt exponential functional forms to measure these spillovers:

$$\Phi(\Lambda) = \exp(\hat{\phi}\Lambda) \tag{103}$$

$$f(\lambda_s^k) = \exp(\hat{f}\lambda_s^k) \tag{104}$$

where Λ is the aggregate level of general R&D efforts:

$$\Lambda = \frac{1}{2}\sum_k \sum_{s=0}^{\infty} \chi_k \mu_s^k (\lambda_s^k + \lambda_{-s}^k) \tag{105}$$

External Calibration. The growth-adjusted interest rate, which is equivalent to the discount rate, is set at $r = 0.02$. Innovation elasticity is set to $\theta = 2$, in line with the microeconometric evidence on innovation (Bloom et al., 2002; Acemoglu et al., 2018). The value of within-sector elasticity of demand, σ, is from De Loecker, Eeckhout and Mongey (2021).

Internal Calibration. The remaining parameters are jointly calibrated using the simulated method of moments, where selected model-generated moments, $\mathbf{m}(\upsilon)$, are compared with their empirical counterparts, $\hat{\mathbf{m}}$, by minimizing the following objective function:

$$\big(\mathbf{m}(\upsilon) - \hat{\mathbf{m}}\big)' \boldsymbol{\Omega}^{-1} \big(\mathbf{m}(\upsilon) - \hat{\mathbf{m}}\big) \quad \text{where } \upsilon = \{\gamma, \kappa^H, \kappa^L, \alpha, \kappa_e, \hat{\phi}, \hat{f}, h\}$$

where $\boldsymbol{\Omega}$ contains weighted squares of the data moments on the main diagonal and zeros elsewhere.[18] While the parameters are calibrated jointly, I provide a discussion on each parameter in relation to the targeted moment it is most closely associated with.

[18]Details about the simulated method of moments are shown in the Appendix.

Model Performance: Targeted Moments. In the model, three parameters are linked to firm-level generality: general R&D cost scale of high type (in R&D efficiency) κ^H, general R&D cost scale of low type κ^L and the efficiency of within-industry knowledge spillovers $\hat{f}$. These parameters are calibrated using three moments related to patent generality: the average patent generality of leaders and followers in high-type industries (SIC 4-digit) and the average patent generality of leaders in low-type industries, based on citation information within a 5-year window following a patent's grant. Specifically, firms with the largest sales in their industries are classified as leaders, while all remaining firms are followers. The innovation step size γ, firm-specific R&D cost scale α, and catch-up rate h are pinned down using three additional moments: an annual aggregate TFP growth rate of 1.66% based on data from Fernald (2015), a median R&D costs to sales ratio of 4.0% and a leadership turnover rate of 23% in high-type industries calculated from the firm sample constructed by linking patent and firm data.[19] The entry cost, κ_e is determined using the average firm entry rate of 11.6% from BDS.[20] Finally, the efficiency of cross-industry knowledge spillovers is pinned down by matching the average ratio of cross-industry to within-industry backward citations of patents.[21] Table 11 documents how the model aligns with the targeted moments.

Model Performance: Untargeted Moments. In addition to the targeted moments, the model also matches three additional moments: average innovation generality of followers and leadership turnover rate in low-type industries and the R&D expenditure to GDP ratio from NCSES. Consistent with the data, the model-implied leadership turnover rate is higher in industries with more efficient general R&D investments.

4.2 Testing the Model's Predictions on Innovation Generality

Using the firm-level dataset for the years 1980 to 2018 constructed in Section 3, I study the patterns of innovation generality within markets. In particular, I test the model predictions from Propositions 1 and 2, that is, whether there exist gaps in innovation generality between leaders and followers and how these gaps vary with the level of market concentration.

[19] The leadership turnover rate is calculated as the fraction of industries with a leading firm different from the one in the last year. The time series of leadership turnover rates for both industry types is provided in the Appendix.

[20] The empirical entry rate is defined as the number of new firms over the number of incumbent firms. Hence the model-implied entry rate equals $\frac{\Phi\lambda_e}{2}$, as in each market, there are two incumbents.

[21] Self-citations are excluded to focus on the knowledge diffusion between firms. Details about the calculation of this ratio in the model are provided in the Appendix.

Table 10: Parameter Values

Parameter	Description	Value	Source
r	Growth-adjusted interest rate	0.02	Standard
σ	Within-sector elasticity of demand	5.75	De Loecker et al. (2021)
θ	Innovation elasticity	2	Bloom et al. (2002)
γ	Innovation step size	1.67	Internal calibration
κ^H	General R&D cost scale, high-type	11.63	Internal calibration
κ^L	General R&D cost scale, low-type	14.06	Internal calibration
α	Firm-specific R&D cost scale	10.59	Internal calibration
κ_e	Entry cost	4.01	Internal calibration
$\hat{f}$	Efficiency of within-industry spillovers	24.90	Internal calibration
$\hat{\phi}$	Efficiency of cross-industry spillovers	45.06	Internal calibration
h	Catch-up rate	0.03	Internal calibration

Note: The first part of the table lists parameters set externally. The second part of the table shows parameters set via internal calibration.

Table 11: Model Fit

	Model	Data
Targeted Moments		
Average innovation generality - Leader, high-type	0.50	0.52
Average innovation generality - Follower, high-type	0.56	0.56
Average innovation generality - Leader, low-type	0.41	0.41
Median R&D to sales ratio	3.7%	4.0%
Aggregate growth rate	1.66%	1.66%
Entry rate	11.6%	11.6%
Leadership turnover, high-type	17%	23%
Average relative backward citations	3.25	3.17
Untargeted Moments		
Average innovation generality - Follower, low-type	0.46	0.46
Leadership turnover, low-type	14%	15%
R&D expenditure to GDP ratio	2.7%	2.5%

Note: Data on innovation generality, R&D to sales ratio, leadership turnover and relative backward citations are calculated from the firm sample constructed by linking USPTO, PatentsView and Compustat. Data on aggregate growth rate are based on Fernald (2015). Data on entry rate are from BDS. R&D to GDP ratio is collected from NCSES. All moments used for calibration are calculated as averages over the 1995–2000 period.

Test of Proposition 1: Leaders have lower innovation generality.

First, I investigate whether there exists a difference in the average innovation generality between leaders and followers from the following OLS estimation:

$$\psi_{f,t} = \underset{(s.e.\ 0.003)}{-\ 0.028} \cdot \mathbb{1}(Leader)_{f,t} + \delta_{j,t} + \epsilon_{f,t} \tag{106}$$

where $\psi_{f,t}$ represents the average patent generality of firm f in year t. $\mathbb{1}(Leader)_{f,t}$ is an indicator function, which equals one if firm f is the leader (i.e., firm with the largest sales) in industry j (with at least two patenting firms) defined according to 4-digit Standard Industrial Classification (SIC) code in year t.[22] $\delta_{j,t}$ captures industry-year fixed effects. I further control for the log of average forward citations at the firm level, where forward citations are counted within a 5-year window following the patent's grant. This accounts for patent quality that may be correlated with the generality score. Holding the leadership or not accounts for about 0.13 of the standard deviation of firm-level innovation generality. This fact is consistent with the model's prediction in Proposition 1.

Note that the estimation results are based on the whole sample from 1980 to 2018. In the US between 1995 and 2000, the average generality gap between leaders and followers was 0.04 in high-type industries and 0.05 in low-type industries, and the parameterized model yields corresponding gaps of 0.06 and 0.05.

Test of Proposition 2: The gap in innovation generality between the two firms varies non-monotonically with market concentration.

Next, based on the firm sample from 1980 t0 2018, I study the relationship between gaps in innovation generality and market concentration, by conducting the following estimation:

$$\psi^L_{j,t} - \psi^F_{j,t} = \underset{(s.e.\ 0.098)}{-\ 0.357} \cdot HHI_{j,t} + \underset{(s.e.\ 0.080)}{0.229} \cdot HHI^2_{j,t} + \delta_j + \delta_t + \epsilon_{j,t} \tag{107}$$

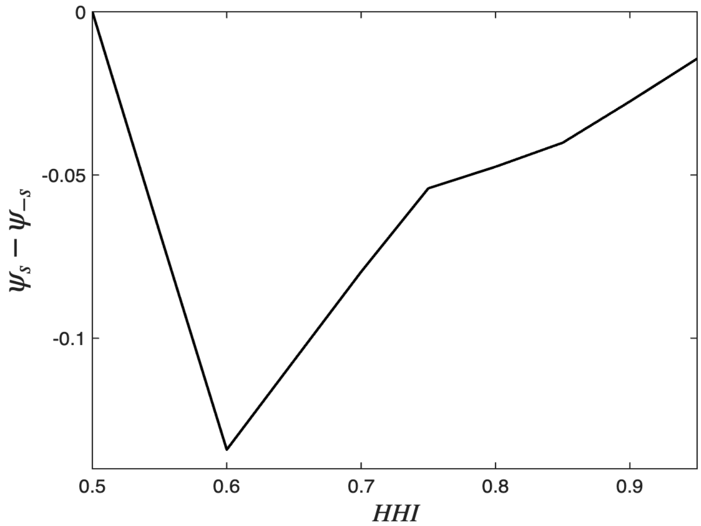
where $\psi^L_{j,t}$ and $\psi^F_{j,t}$ represent average patent generality of the leader and follower in industry j in year t, respectively. $HHI_{j,t}$ is Herfindahl–Hirschman index (HHI), defined as the sum of the squared market shares of all firms in industry j in year t, which is a standard measure of market concentration.[23] The term $HHI^2_{j,t}$ is the square of $HHI_{j,t}$. δ_j and δ_t capture industry fixed effects and year fixed effects, respectively. I further control for the total number of firms in the industry. This estimation indicates that generality

[22] This definition of market leader is common in the literature (see e.g., Kroen, Liu, Mian and Sufi, 2021; Olmstead-Rumsey, 2025). The results are robust if firms whose sales are above the top 5 percentile are considered as leaders. See Appendix.

[23] For industries with multiple followers, I use the median value of the innovation generality among followers. The results remain robust if the mean value is used. See Appendix.

gaps follow a U-shaped pattern with respect to market concentration, consistent with the theoretical prediction in Proposition 2.[24] Figure 18 illustrates the U-shaped pattern of the parameterized model. For each level of market concentration, the figure shows the weighted average of the generality gap across the two industry types.

Figure 18: Model-implied Generality Gaps and Market Concentration



Note: The figure plots the model-implied relationship between gaps in innovation generality, $(\psi_s - \psi_{-s})$, and market concentration, measured by HHI.

To conclude, these two estimations show that there is a discrepancy in innovation generality, defined as the average patent generality, between leaders and followers. Leaders prefer innovations with lower generality, while followers tend to produce more general innovations, consistent with Proposition 1. However, the gaps in innovation generality between the leader and the follower have a U-shaped relationship with the level of market concentration, in line with Proposition 2.

4.3 Model Validation: Within-industry Spillovers

The central mechanism behind firms' strategic innovation decisions is within-industry general knowledge spillovers. This mechanism has two key elements: (1) general knowledge flows from leaders to followers, and (2) absorbing these spillovers requires followers themselves to engage in general knowledge production. The model therefore predicts that when within-industry spillovers are more efficient, leaders will shift toward firm-specific innovations to protect their advantage, while followers will invest more in general innovations to absorb external general knowledge and catch up. Conversely, when within-industry spillovers weaken, leaders become more willing to produce general knowledge, while followers are discouraged from doing so.

[24]The bin plot of this empirical relationship is shown in Figure 43 in the Appendix.

To validate this mechanism, I draw on a quasi-natural experiment that leverages variation in state-level enforcement of non-compete agreements, which serves as a proxy for the efficiency of within-industry spillovers. The logic behind this is that non-compete agreements restrict workers from competing against their former employer within a certain period, thereby directly affecting the mobility of R&D workers—a key channel for knowledge diffusion (Jaffe et al., 1993; Almeida and Kogut, 1999; Singh and Agrawal, 2011; Stoyanov and Zubanov, 2012; Liu, 2023). Stronger enforcement of non-competes limits labor mobility and thereby reduces the efficiency of within-industry knowledge spillovers.

State-level Enforceability of Non-compete Agreements. I use a state-year index of non-compete enforceability, first developed by Bishara (2011) and later extended by Marx (2022), which incorporates judicial and legislative decisions affecting the enforceability of non-compete agreements for all workers. The state-level changes in non-compete enforceability can be treated as plausibly exogenous legal shocks (see, e.g., Marx, Strumsky and Fleming, 2009; Samila and Sorenson, 2011). Following Reinmuth and Rockall (2025) and Ma, Wang and Wu (2025), I rely on the normalized version of this index to measure the strength of enforcement across states, to ensure the interpretability of regression coefficients. To capture major policy shifts, I construct a binary treatment variable, NCA, which indicates whether a state has recently experienced a significant change in enforceability—defined as a year-over-year change in the index greater than 0.02—following the clean treatment approach in Reinmuth and Rockall (2025) and Ma et al. (2025).[25] The variable equals 1 if the change represents a strengthening of enforcement, −1 if it represents a weakening, and 0 otherwise. Using this definition, I identify 16 major changes between 1991 and 2014, of which 11 involved stronger enforcement.[26]

Estimation. To examine how firms respond to changes in the efficiency of within-industry general knowledge spillovers, proxied by shifts in non-compete agreement enforceability, I estimate the following difference-in-differences specification:

$$\psi_{f,t} = \beta_N NCA_{s,t} + \beta_L \mathbb{1}(Leader)_{f,t} + \beta_{N,L} \left[NCA_{s,t} \times \mathbb{1}(Leader)_{f,\tau} \right] + \delta_{j,t} + \delta_s + \epsilon_{f,t} \quad (108)$$

where $\psi_{f,t}$ is the average patent generality of firm f in year t. The variable $NCA_{s,t}$ indicates the direction of the last major reform in the enforceability of non-compete

[25]Note that the results remain robust when a significant change is alternatively defined as a larger or smaller shift in the index. See Appendix.

[26]Data on state-level enforceability of non-compete agreements are available for the period 1991–2014. Accordingly, the combined firm sample is restricted to this period.

agreements in the state s, where firm f's headquarters is located. It equals 1 if the last reform reflects stronger enforcement, -1 if it reflects weaker enforcement, and 0 otherwise. The indicator $\mathbb{1}(Leader)_{f,t}$ equals one if firm f is an industry leader in year t, capturing contemporaneous differences in innovation generality between leaders and followers. The interaction term $\left[NCA_{s,t} \times \mathbb{1}(Leader)_{f,\tau}\right]$ identifies whether firms that were industry leaders at the time of the reform (i.e., τ) respond differently to changes in non-compete enforceability.[27] Finally, $\delta_{j,t}$ and δ_s represent industry-year and state fixed effects, respectively, which absorb time-varying industry-specific shocks and time-invariant heterogeneity across states. The key coefficients of interest are β_N and $\beta_{N,L}$: β_N captures the effect of changes in non-compete enforceability on followers; the coefficient $\beta_{N,L}$ captures the differential effect of changes in non-compete enforceability on leaders relative to followers.

Table 12 reports the estimated coefficients, with standard errors clustered at the state level. Importantly, β_N is negative and statistically significant, and $\beta_{N,L}$ is positive and statistically significant. Moreover, the magnitude of $\beta_{N,L}$ is larger than that of β_N (i.e., $\beta_{N,L} + \beta_N > 0$). This indicates that when non-compete enforcement strengthens, leaders shift toward more general innovations, whereas followers tilt toward more firm-specific innovations. These findings align with the model's logic of within-industry knowledge spillovers: when spillovers weaken, leaders can safely produce more general innovations without fear of being overtaken, while followers lose incentives to produce general innovations as a means of absorbing external knowledge and catching up.

Table 12: Within-industry General Knowledge Spillovers

β_N	-0.016^{**}
	(0.008)
$\beta_{N,L}$	0.037^{***}
	(0.008)
Industry-Year FE	✓
State FE	✓
R-squared	0.454
Observations	23,764

Standard errors are clustered at the state level.
*** p<0.01, ** p<0.05, * p<0.1

Note: This table shows the results from estimation (108).

[27] As a robustness check, I reconstruct the leadership indicator in two alternative ways: (1) using predetermined leader status measured one or two periods before the treatment year, to mitigate potential post-treatment bias; and (2) using a measure of persistent leadership, defined as maintaining leader status from the time of the reform (i.e., τ) through period t. See Appendix for details.

In the parameterized model, reducing the coefficient for within-industry knowledge spillover efficiency ($\hat{f}$) from 24.9 to 8.4 leads to a rise in the average innovation generality of leaders across the two industry types from 0.452 to 0.472 (an increase of 0.020). Meanwhile, the average innovation generality of followers declines from 0.508 to 0.492 (a decrease of 0.016). These patterns align closely with the estimation results reported in column (3) of Table 12, where leaders' innovation generality increases by $(\beta_{N,L} + \beta_N)$ (0.021) to a stricter enforcement of non-compete agreements, while followers' innovation generality falls by β_N (0.016).

4.4 Firms' Innovation Decisions and Within-industry Spillovers

Having validated the model's within-industry general knowledge spillovers, I next examine how leaders and followers adjust their innovation decisions in response to these spillovers, based on the parametrized model.

Without loss of generality, I focus on the high-type industries. Figure 19 shows the general and firm-specific R&D of the leader and follower in each state s, the stationary distribution, and the extent of within-industry knowledge spillovers, which is measured by the additional growth that followers gain through within-industry knowledge spillovers, and can be expressed as:

$$\Phi\big[f(\lambda_s^H) - 1\big]\lambda_{-s}^H = \Phi\big[\exp(\hat{f}\lambda_s^H) - 1\big]\lambda_{-s}^H \tag{109}$$

$$\approx \Phi\hat{f}\lambda_s^H\lambda_{-s}^H \tag{110}$$

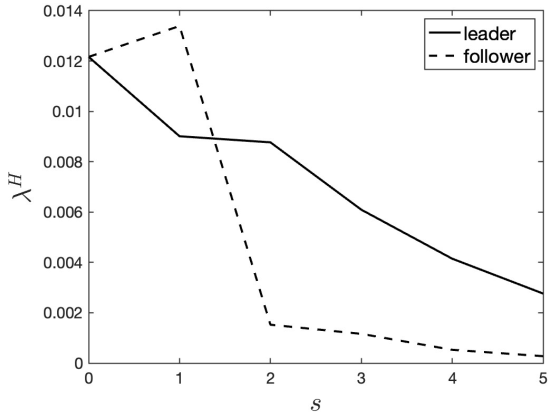
where $\Phi(\Lambda) = \exp(\hat{\phi}\Lambda)$ measures the external effects of cross-industry general knowledge, $\hat{f}$ governs the efficiency of within-industry knowledge spillovers, λ_s^H and λ_{-s}^H are the general R&D efforts of the leader and the follower, respectively.

As predicted in Proposition 8, when the distance between leaders and followers is small but positive, followers have strong incentives to pursue general innovations to absorb the external general knowledge produced by leaders, as the potential profit gains are substantial. Conversely, leaders prioritize firm-specific knowledge to mitigate the risk of being overtaken. However, as the distance widens, these incentives weaken.

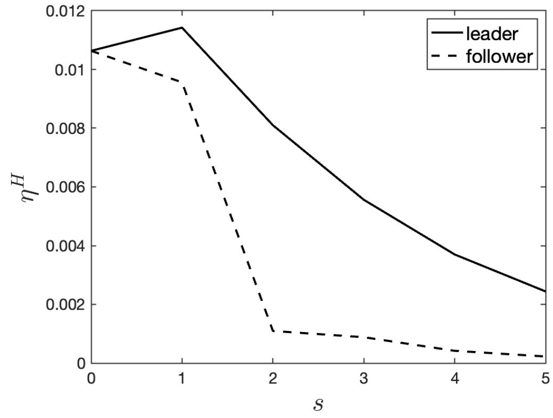
In state $s = 0$, the within-industry spillovers is muted by definition (i.e., $f(0) = 1$). At $s = 1$, the extent of within-industry knowledge spillovers reaches its peak, as the follower invests heavily in general R&D to absorb the external general knowledge and improve its productivity z_{-s}. As the distance s increases, the probability of the follower overtaking the leader decreases. Therefore, the leader faces less competition and reduces its innovation efforts, allocating resources more evenly between general and firm-specific R&D. Meanwhile, as the follower falls further behind, its probability of catching up

Figure 19: R&D Efforts, Spillovers and Distribution (High-type Industries)

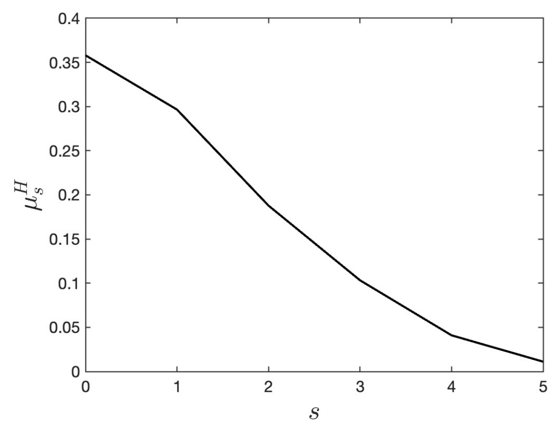
(a) General R&D (λ^H)



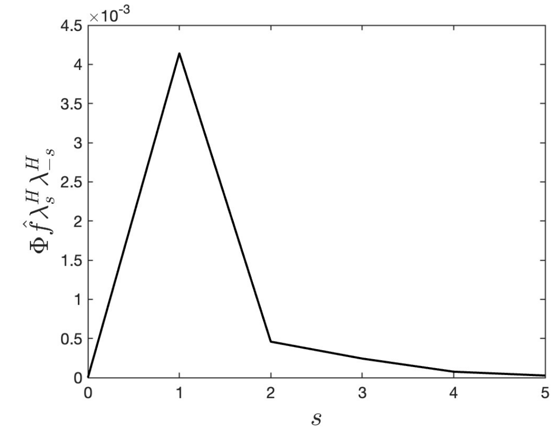
(b) Firm-specific R&D (η^H)



(c) Stationary Distribution (μ_s^H)



(d) The Extent of Within-industry Spillovers



Note: Panel (a) plots general R&D efforts of the leader (solid line) and the follower (dashed line) in each state. Panel (b) plots the corresponding firm-specific R&D efforts. Panel (c) plots stationary distribution of productivity gap s. Panel (d) plots the extent of within-industry knowledge spillovers.

declines, weakening its incentive to invest in R&D.

4.5 Market Concentration and Innovation Generality

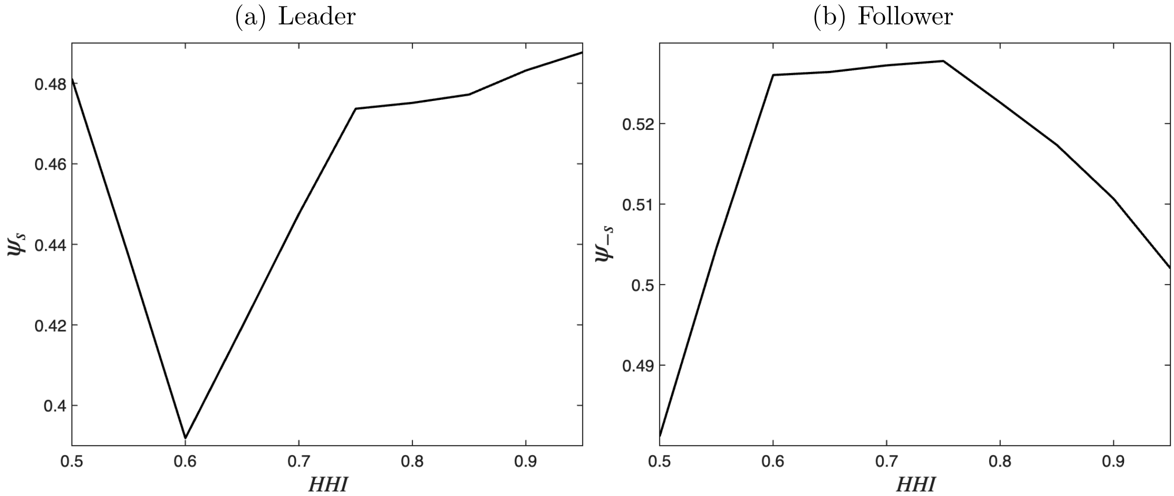
After studying firms' innovation decisions, I turn to the link between innovation generality and market concentration. Note that the level of market concentration in the model has a one-to-one relationship with the distance between the leader and the follower.

$$HHI_s = \big(\frac{\rho_s^{1-\sigma}}{\rho_s^{1-\sigma}+1}\big)^2 + \big(\frac{1}{\rho_s^{1-\sigma}+1}\big)^2$$

with ρ_s the price ratio between the leader and the follower, which satisfies $\rho_s^\sigma = \gamma^{-s}\frac{\sigma\rho_s^{\sigma-1}+1}{\sigma+\rho_s^{\sigma-1}}$.

As predicted in Proposition 2, the generality of leaders' and followers' innovations varies systematically with concentration. In the parametrized model, leaders exhibit a U-shaped relationship between innovation generality and concentration, while followers display a hump-shaped pattern, as shown in Figure 20.

Figure 20: Innovation Generality and Market Concentration



Note: Panel (a) plots the relationship between the leader's innovation generality, ψ_s, with level of market concentration, measured by HHI. Panel (b) plots that of the follower. Both figures plot the weighted average of innovation generality across the two industry types.

To compare these patterns in the data, I investigate the empirical relationship between market concentration and innovation generality of leaders and followers, separately, following estimation (107) from the test of Proposition 2:

$$\psi^L_{j,t} = \underset{(s.e.\ 0.070)}{-\ 0.194} \cdot HHI_{j,t} + \underset{(s.e.\ 0.057)}{0.137} \cdot HHI^2_{j,t} + \delta_j + \delta_t + \epsilon_{j,t} \tag{111}$$

$$\psi^F_{j,t} = \underset{(s.e.\ 0.071)}{0.164} \cdot HHI_{j,t} - \underset{(s.e.\ 0.058)}{0.092} \cdot HHI^2_{j,t} + \delta_j + \delta_t + \epsilon_{j,t} \tag{112}$$

where $\psi^L_{j,t}$ and $\psi^F_{j,t}$ represent average patent generality of the leader and follower. The estimation results are qualitatively consistent with the model and reinforce its mechanism: leaders limit general knowledge production to preserve their advantage, while followers rely on it to absorb within-industry spillovers and narrow the technological gap.[28] These incentives are strongest when the leader–follower distance is small (i.e., $s = 1$).

[28]The results are robust if further controlling for the log of average forward citations at the firm level. See Appendix.

5 Policy Implications and Secular Trends

In this section, I study the policy implications based on the parameterized model and secular trends of knowledge generality. I begin with the comparative statics of aggregate growth, which show that growth is more responsive to general R&D costs than to firm-specific R&D costs, suggesting that subsidies for general R&D are more effective.[29] Then I evaluate the impact of general R&D subsidies, and consider the effects of improving the efficiency of general knowledge spillovers.[30] Finally, I document the secular trends in U.S. aggregate growth and innovation generality, and discuss how they are related.

5.1 Comparative Statics of Aggregate Growth

Before discussing the policy implications, I provide comparative statics on the key factors that impact growth. Figure 21 shows how the aggregate growth rate responds to a reduction in R&D costs for all firms across all industries, or an increase in the efficiency of general knowledge spillovers.

Decrease in R&D Costs. Panels (a) and (b) plot the growth elasticity with respect to (a decline in) general and firm-specific R&D cost scale, respectively. From panel (a), we observe that lowering the general R&D cost scale (i.e., κ_k) across all industries consistently increases economic growth. When incumbents engage more in general innovations, they experience faster firm-level growth. In addition, this increases the level of aggregate general knowledge (i.e., Λ), which further reinforces cross-industry spillovers through a feedback loop (i.e., $\Phi(\Lambda)$).

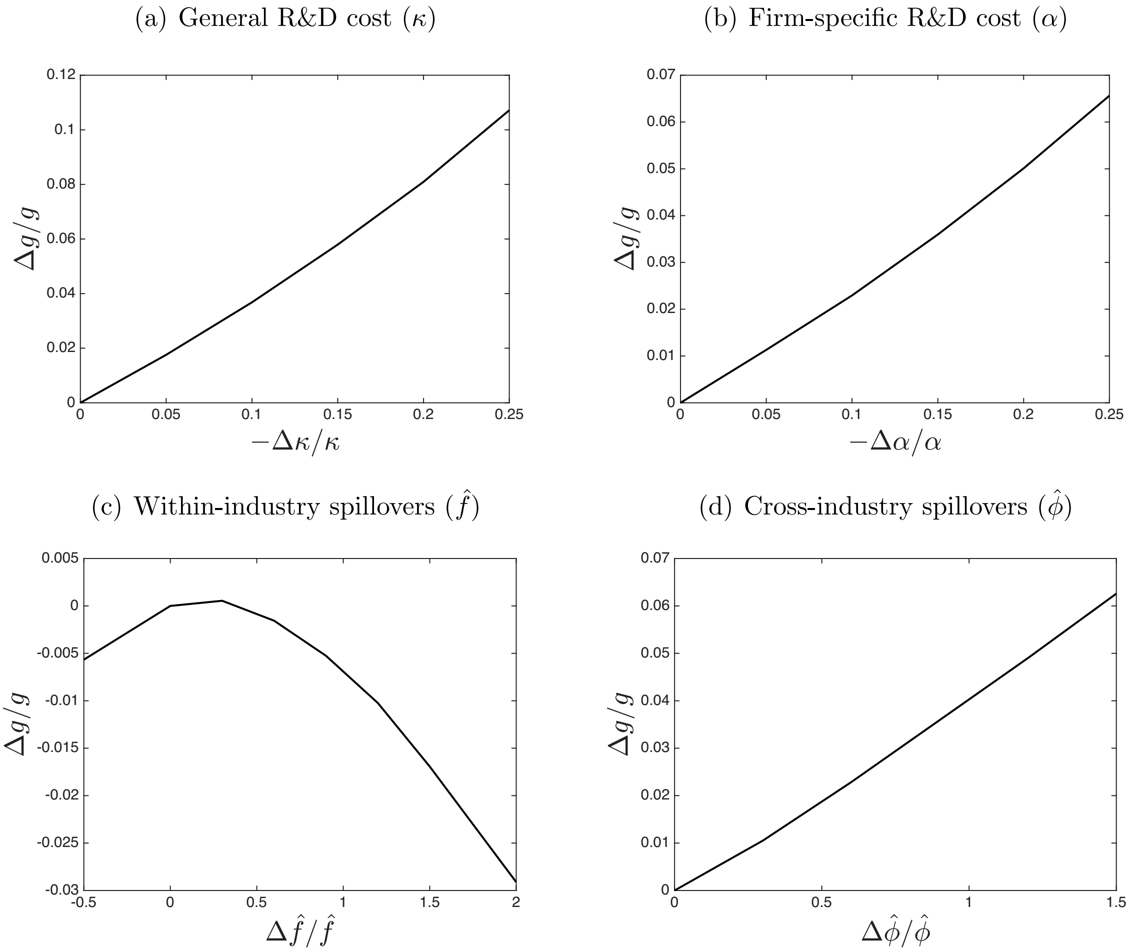
Reducing firm-specific R&D cost scale (i.e., α) also leads to higher aggregate growth but by a much lower magnitude, as shown in panel (b). While incumbents intensify their firm-specific R&D efforts, the benefits are partially offset by reduced cross-industry knowledge spillovers, due to a decline in aggregate general knowledge. Quantitatively, the growth elasticity with respect to general R&D costs is nearly twice that of firm-specific R&D costs.

Improvement in the efficiency of general knowledge spillovers. In traditional models without knowledge generality, improving the efficiency of knowledge spillovers, represented by a higher constant catch-up rate, directly leads to faster aggregate growth

[29]Note that along the BGP, total output, total income, economy-wide profits and wage for production workers all grow at the same rate, g.

[30]The analysis focuses on inefficiencies from knowledge spillovers rather than distortions from firms' pricing. In this setting, policies that maximize aggregate growth also maximize welfare, since firms' innovation expenditures ultimately return to households through R&D employment.

Figure 21: Growth Elasticity w.r.t. R&D Costs and Efficiency of Spillovers



Note: Panel (a) and (b) plot the elasticity of growth with respect to the general and firm-specific R&D cost scale. Panel (c) and (d) plot the elasticity of growth with respect to the efficiency of knowledge spillovers within- and cross-industry. For costs, changes in growth are shown with a decrease in costs. For spillovers, changes in growth are shown with an increase in the efficiency of spillovers.

(see e.g., Liu et al., 2022; Akcigit and Ates, 2023). However, when innovation generality is taken into account, the effects of more efficient within-industry knowledge spillovers may turn negative for growth, as firms adjust their innovation strategies in response to the extent of knowledge diffusion.

Panels (c) and (d) plot the growth elasticity with respect to (an increase in) the efficiency of within- and cross-industry general knowledge spillovers. Panel (c) suggests that the relationship between aggregate growth and the efficiency of within-industry spillovers (i.e., $\hat{f}$) follows an inverted U-shape. When within-industry spillovers become highly efficient, aggregate growth may decline, as leaders are strongly discouraged from investing in general innovations. This reduction offsets followers' expansion of general R&D, reducing the aggregate level of general knowledge and limiting the extent of cross-industry spillovers.

In contrast, enhancing the efficiency of cross-industry knowledge spillovers (i.e., $\hat{\phi}$) leads to higher growth, as shown in panel (d). This is because it encourages incumbents to conduct general innovations, which improves the extents of both within- and cross-industry knowledge spillovers.

5.2 Policy Implications

The comparative statics indicate that growth is more elastic with respect to general R&D costs than to firm-specific R&D costs, implying that subsidies for general R&D are more effective.

5.2.1 General R&D Subsidies for High-type Industries

The first policy experiment relates to R&D subsidies and follows a traditional approach, assuming a fixed corporate tax rate δ for all firms across all industries. I begin with the case where subsidies are directed toward general innovations in high-type industries, and then compare it with the scenario of targeting general innovations in low-type industries.

Denote the implied R&D subsidies ν_L^H for leaders, ν_F^H for followers and ν_0^H for neck-and-neck firms. Then the HJB equations (92)-(94) can be rewritten as:

$$\begin{aligned}
rv_s^H &= \max_{\eta_s^H,\lambda_s^H\geq 0}\Big\{(1-\delta)\pi_s - c_\eta(\eta_s^H) - (1-\nu_L^H)c_\lambda^H(\lambda_s^H) \\
&\quad +[\eta_{-s}^H + \Phi f(\lambda_s^H)\lambda_{-s}^H + h](v_{s-1}^H - v_s^H) + (\eta_s^H + \Phi\lambda_s^H)(v_{s+1}^H - v_s^H) - \tau^H v_s^H\Big\} \\
rv_{-s}^H &= \max_{\eta_{-s}^H,\lambda_{-s}^H\geq 0}\Big\{(1-\delta)\pi_{-s} - c_\eta(\eta_{-s}^H) - (1-\nu_F^H)c_\lambda^H(\lambda_{-s}^H) \\
&\quad +[\eta_{-s}^H + \Phi f(\lambda_s^H)\lambda_{-s}^H + h](v_{-s+1}^H - v_{-s}^H) + (\eta_s^H + \Phi\lambda_s^H)(v_{-s-1}^H - v_{-s}^H) - \tau^H v_{-s}^H\Big\} \\
rv_0^H &= \max_{\eta_0^H,\lambda_0^H\geq 0}\Big\{(1-\delta)\pi_0 - c_\eta(\eta_0^H) - (1-\nu_0^H)c_\lambda^H(\lambda_0^H) \\
&\quad +(\eta_{-0}^H + \Phi\lambda_{-0}^H)(v_{-1}^H - v_0^H) + (\eta_0^H + \Phi\lambda_0^H)(v_1^H - v_0^H) - \tau^H v_0^H\Big\}
\end{aligned}$$

subject to a budget constraint:

$$\sum_{k=\{H,L\}}\sum_{s=0}^{\infty}\delta(\pi_s+\pi_{-s})\mu_s^k = 2\nu_0^H c_\lambda^H(\lambda_0^H)\mu_0^H + \sum_{s=1}^{\infty}\left[\nu_L^H c_\lambda^H(\lambda_s^H) + \nu_F^H c_\lambda^H(\lambda_{-s}^H)\right]\mu_s^H \tag{113}$$

This budget constraint says that the overall corporate taxes collected are used to cover all the general R&D subsidies among high-type industries. I consider two schemes: subsidizing leaders (and neck-and-neck firms) only, or subsidizing followers (and neck-and-neck firms) only. Table 13 shows the results for each scenario.

Surprisingly, subsidizing general R&D for leaders is almost always more beneficial than

Table 13: General R&D Subsidies for High-type Industries

Tax rate	δ	0.1	0.15	0.2	0.25
Scheme 1: Subsidizing Leaders					
Growth rate	g	1.76%	1.82%	1.88%	1.95%
Average Generality - Leaders	ψ_s^H	0.65	0.71	0.77	0.83
Average Generality - Followers	ψ_{-s}^H	0.60	0.61	0.62	0.63
Scheme 2: Subsidizing Followers					
Growth rate	g	1.73%	1.77%	1.80%	1.80%
Average Generality - Leaders	ψ_s^H	0.48	0.47	0.47	0.47
Average Generality - Followers	ψ_{-s}^H	0.67	0.71	0.74	0.79

Note: The first row shows the fixed tax rates for all incumbent firms. Scheme 1 reports counterfactuals when only leaders and neck-and-neck firms in high-type industries are subsidized. The calculation of ψ_{-s}^H excludes the innovation generality of neck-and-neck firms. Scheme 2 reports counterfactuals when only followers and neck-and-neck firms in high-type industries are subsidized. The calculation of ψ_s^H excludes the innovation generality of neck-and-neck firms.

supporting followers, a result that differs from the policy implications of the traditional leader–follower framework (see e.g., Liu et al., 2022; Akcigit and Ates, 2023). As shown in the table, implementing a 15% tax rate and using the revenue to subsidize general innovations by leaders in high-type industries raises the aggregate growth rate from 1.66% to 1.82%. In contrast, directing the subsidies to followers results in only two-thirds of that growth improvement. This is because when followers receive general R&D subsidies, they focus more on producing general knowledge, which aids their catch-up process. Anticipating this, leaders reduce their general R&D efforts to maintain their dominance. As a result, the overall positive impact of general R&D subsidies on aggregate general knowledge is mitigated, resulting in limited improvement in the cross-industry spillovers. Conversely, subsidizing leaders' general R&D encourages them to invest more in general innovations, which, in turn, motivates followers to enhance their own R&D efforts to absorb the increasing external general knowledge. This feedback effect further pushes up the aggregate level of general R&D efforts, and benefits all incumbents through cross-industry spillovers (i.e., Φ).

Comparison with General R&D Subsidies in Low-type Industries. For comparison, I provide a discussion when general innovations in low-type industries are subsidized instead. Table 14 provides the results. The first part shows the aggregate growth rate, average innovation generality when general innovations of leaders and neck-and-neck firms in low-type industries are subsidized. The second part shows the results when only followers and neck-and-neck firms in low-type industries are subsidized.

Subsidizing general R&D innovations in low-type industries still promotes growth, and targeting leaders is more effective than targeting followers—consistent with the results for high-type industries. However, the growth gains are substantially smaller. For instance, implementing a 15% tax rate and subsidizing leaders' general innovations in low-type industries increases the aggregate growth rate from 1.66% to 1.75%, which is half the improvement observed in the high-type case. This is because, for the same level of subsidies, incumbents in low-type industries produce less general knowledge and create weaker cross-industry spillovers.

Table 14: General R&D Subsidies for Low-type Industries

Tax rate	δ	0.1	0.15	0.2	0.25
Scheme 1: Subsidizing Leaders					
Growth rate	g	1.72%	1.75%	1.78%	1.81%
Average Generality - Leaders	ψ_s^L	0.54	0.60	0.65	0.70
Average Generality - Followers	ψ_{-s}^L	0.50	0.51	0.51	0.52
Scheme 2: Subsidizing Followers					
Growth rate	g	1.71%	1.72%	1.74%	1.75%
Average Generality - Leaders	ψ_s^L	0.39	0.39	0.39	0.38
Average Generality - Followers	ψ_{-s}^L	0.57	0.61	0.64	0.67

Note: The first row shows the fixed tax rates for all incumbent firms. Scheme 1 reports counterfactuals when only leaders and neck-and-neck firms in low-type industries are subsidized. The calculation of ψ_{-s}^L excludes the innovation generality of neck-and-neck firms. Scheme 2 reports counterfactuals when only followers and neck-and-neck firms in low-type industries are subsidized. The calculation of ψ_s^L excludes the innovation generality of neck-and-neck firms.

To conclude, policies aimed at targeting general R&D of leaders in high-type industries can maximize knowledge spillovers and stimulate aggregate growth efficiently. This policy implication differs from the traditional leader–follower framework in the literature (see, e.g., Liu et al., 2022; Akcigit and Ates, 2023), where subsidizing followers is typically regarded as the optimal policy. Supporting followers remains valuable, as they are often self-motivated to engage in general innovations and require less monitoring. However, directing subsidies toward general innovations in leading firms can foster aggregate growth more efficiently, as it strengthens their incentives to pursue knowledge with broader applicability and amplifies both within- and cross-industry spillovers. At the same time, because leaders tend to focus on firm-specific innovations to protect their market position, careful oversight of subsidy allocation is required. A more detailed discussion of the policy design is provided in section 5.2.3.

5.2.2 Improving the Efficiency of Cross-industry Spillovers

The comparative statics suggest that improving the efficiency of within-industry general knowledge spillovers does not necessarily increase growth, since it reduces leaders' incentives to undertake general innovations, which limits the spillovers available to firms in other industries. This implies that weak non-compete agreements do not necessarily improve growth. If inventors with general knowledge can easily move to follower firms, industry leaders may be discouraged from pursuing such innovations in the first place.

Instead, policies that improve the efficiency of cross-industry general knowledge spillovers can foster aggregate growth. For example, interdisciplinary research centers that promote collaboration across sectors can help firms adapt advances from other fields into new products, production methods, and business models within their own industries.

5.2.3 Discussion

Since the empirical measure of patent generality depends on forward citations, it has limited practical use for policies like R&D subsidy allocation. Instead, we need measures available prior to patent grant that can predict the generality of a patent.

Patent Originality. The originality measure is built upon patents' backward citations, and is formally defined as:

$$o_i = 1 - \sum_j c_{ij}^2$$

where c_{ij} is the fraction of citations that patent i makes that belong to patent class j. A high originality score indicates that the patent cites previous patents in a wide set of technologies, whereas a low score implies a narrow range of fields the patent is built on (Hall, Jaffe and Trajtenberg, 2001; Babina, He, Howell, Perlman and Staudt, 2023).

Importantly, patent generality is positively and significantly associated with its originality score:

$$\psi_{i,t} = \underset{(s.e.\ 0.001)}{0.223} \cdot o_{i,t} + \delta_f + \delta_t + \epsilon_{i,t} \tag{114}$$

where $\psi_{i,t}$ is patent i's generality score, and $o_{i,t}$ is its originality score, with t its application year. δ_f and δ_t represent firm fixed effects and application year fixed effects, respectively.

Therefore, patent originality, which is based on backward citations, can serve as a predictor of how general the knowledge is. However, firms may strategically manipulate backward citations to inflate the perceived originality of their patents. To discourage such behavior, a penalty system can be introduced. For instance, if a patent is later found to have deliberately excluded key citations or include irrelevant citations, a reduction in subsidy for general innovations can be applied to the same firm.

Skill Complementarity. In addition to patent originality scores, human capital can also serve as a predictor of knowledge generality. For example, research teams composed of inventors from diverse backgrounds may generate more general knowledge due to skill complementarity.

Other criteria, such as a high self-cite rate and long patent claims possibly imply that the patent is less general and of narrower scope.[31]

Prize Instead of Subsidy. Admittedly, identifying general innovations before a patent is granted can be challenging. An alternative policy instrument—mathematically similar to an R&D subsidy but differing in timing—is a prize system to encourage general innovations. Recent work shows that cash prizes can complement subsidies by steering innovation toward socially valuable outcomes (see, e.g., Che et al., 2021; Graff Zivin and Lyons, 2021). Crucially, unlike subsidies, which are provided upfront, prizes are granted after the innovation is developed, thus reducing the opportunity for firms to strategically manipulate backward citations to influence funding decisions. For instance, the King's Award for Innovation in the United Kingdom recognizes firms that deliver outstanding technological advancements. A prize system that rewards leading firms for conducting general innovations can enhance both within- and cross-industry knowledge spillovers, thereby stimulating aggregate growth.

5.3 (General) Ideas Become Harder to Find

This part quantifies the impact of the post-2010 decline in general R&D efficiency on aggregate growth and business dynamism.

Trends in Average Patent Generality and TFP Growth. In their seminal paper, Bloom, Jones, Van Reenen and Webb (2020) show the trends in rising research effort and declining research productivity (i.e., ideas are getting harder to find). Building on their findings, I provide evidence that general ideas, in particular, are becoming harder to find. Figure 22 shows the trend in average patent generality by industry type and aggregate productivity growth in the US. Panel (a) presents the generality measure based on citation information within 5-year periods following a patent's grant. Panel (b) plots the annual TFP growth based on Fernald (2015), and is smoothed using an HP filter with an annual

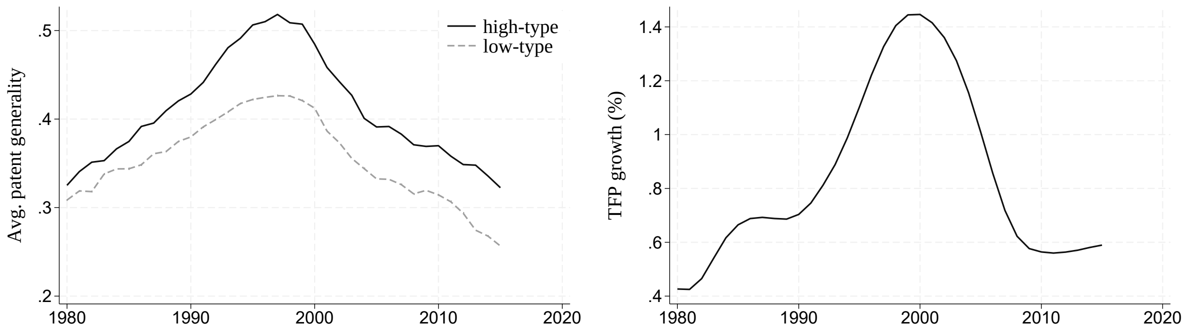
[31] Akcigit and Ates (2023) use self-cite rate and length of patent claims to infer the strategic use of patents to limit the scope of spillovers to competitors.

smoothing parameter of 100. The average patent generality reached its peak prior to 2000, followed by a consistent decline.[32] The growth of TFP shows a similar trend.

This observed trend in aggregate patent generality is broadly shared across most major sectors (see Figure 44 in Appendix B.4). A similar pattern also appears in the average leadership turnover rate (see Figure 47 in Appendix B.7).

Figure 22: Average Patent Generality and Aggregate Growth

(a) Average Patent Generality (b) TFP Growth



Note: Panel (a) plots the average patent generality in the patent sample from 1980 to 2015. The black solid line and the gray dashed line show the average patent generality in high- and low-type industries, respectively. Panel (b) plots the annual productivity growth using Fernald (2015). The plot is smoothed using an HP filter with an annual smoothing parameter of 100.

Calibration to Post-2010 Data. During the 2010-2015 period, the average innovation generality of leaders and followers declined to 0.34 and 0.38 in high-type industries, and to 0.30 and 0.33 in low-type industries. Meanwhile, the average TFP growth rate dropped to 0.98%. To align the model with these post-2010 statistics, I apply the simulated method of moments, as in the previous calibration. The parameters adjusted include the general R&D cost scale in high- and low-type industries (κ^H and κ^L), the firm-specific R&D cost scale (α), the efficiency of within- and cross-industry spillovers ($\hat{f}$ and $\hat{\phi}$), as listed in Table 15. In particular, there is a reduction in the efficiency of both general and firm-specific R&D investments, in line with Bloom et al. (2020).

The first half of Table 16 lists the targeted moments, while the second half reports the untargeted ones. The lower efficiency of general R&D investments—especially in high-type industries—reduces incumbents' incentives to produce general knowledge. The resulting decline in aggregate general knowledge discourages entry and makes it harder for followers to catch up. Together, these forces generate a lower leadership turnover rate. Since the efficiency decline is more pronounced in high-type industries, the model

[32]Note that this pattern is not driven by changes in the number of forward citations or in overall patenting activity. The corresponding trends appear in Figure 45 in Appendix B.5 and Figure 46 in Appendix B.6.

predicts a larger reduction in turnover there, consistent with the data. Moreover, this decline in leadership turnover arises endogenously from the reduced efficiency of R&D investments faced by both leaders and followers, rather than from a decrease in the fixed catch-up rate (h).

Table 15: Parameter Values: Post-2010

Parameter	Description	Value	Source
κ^H	General R&D cost scale, high-type	21.09	Internal calibration
κ^L	General R&D cost scale, low-type	23.32	Internal calibration
α	Firm-specific R&D cost scale	14.52	Internal calibration
$\hat{f}$	Efficiency of within-industry spillovers	28.50	Internal calibration
$\hat{\phi}$	Efficiency of cross-industry spillovers	55.04	Internal calibration

Note: The table shows parameters that match the post-2010 US economy and are set via internal calibration.

Table 16: Model Fit: Post-2010

	Model	Data
Targeted Moments		
Average innovation generality - Leader, high type	0.35	0.34
Average innovation generality - Follower, high type	0.37	0.38
Average innovation generality - Leader, low-type	0.31	0.30
Aggregate growth rate	0.98%	0.98%
Leadership turnover, high-type	12%	11%
Untargeted Moments		
Average innovation generality - Follower, low-type	0.33	0.33
Entry rate	8.80%	9.86%
Leadership turnover, low-type	11%	10%
Average relative backward citations	3.73	4.27

Note: Data on innovation generality, leadership turnover rate and relative backward citations are calculated from the firm sample constructed by linking USPTO, PatentsView and Compustat. Data on aggregate growth rate are based on Fernald (2015). Data on entry rate are from BDS. All moments used for calibration are calculated as averages over the 2010–2015 period.

Decomposition of the Decline in Aggregate Growth. Next, I analyze the channels underlying the post-2010 slowdown in U.S. aggregate growth. Table 17 presents the decomposition results. Column (1) reports the decomposition method, with channels added sequentially. Consistent with the parameter adjustments in Table 15, I consider three channels: less efficient general R&D in high-type industries (an increase in κ^H), less efficient general R&D in low-type industries (an increase in κ^L), less efficient firm-specific R&D (an increase in α). Column (2) presents the model-implied growth rate, and Column (3) reports the contribution of the accumulated channels to the change in aggregate growth. The decline in the overall efficiency of general R&D accounts for half of the growth slowdown, with around 60% of this effect driven by less efficient general R&D in high-type industries.

Table 17: Decomposition of the Decline in Aggregate Growth

	Implied g	Contribution
Less eff general R&D - high-type only	1.46%	30%
Less eff general R&D - all industries	1.32%	51%
Less eff general + firm-specific R&D	0.98%	100%

Note: The first column shows the decomposition method. The second column reports the implied aggregate growth rate. The third column shows the accumulation of the contribution to the change in aggregate growth.

Innovation Generality and Business Dynamism. As predicted by the model, the efficiency of general R&D is closely tied to business dynamism through within- and cross-industry spillovers, as well as entrants' reliance on general knowledge. Since 2000, the U.S. economy has experienced a marked decline in entry rate and leadership turnover (Bessen, Denk, Kim and Righi, 2020; Olmstead-Rumsey, 2025). This decline, according to the calibrated model, is primarily driven by a reduction in aggregate general knowledge production, which hinders diffusion from incumbents to entrants and, in turn, lowers the endogenous entry rate. At the same time, lower general R&D efficiency reduces leadership turnover in both types of industries: leaders are less likely to be replaced by entrants, and followers struggle to catch up. Table 18 summarizes how the decline in general R&D efficiency across both industry types contributes to the reduction in business dynamism between the pre-2000 and post-2010 periods. The first column lists the dimensions of business dynamism, and the second column reports the model-implied values given the lower general R&D efficiency in the post-2010 period. The third column reports its contribution, by comparing the baseline model (calibrated to pre-2000 U.S. economy) with a counterfactual in which only the general R&D cost scale is updated to its post-2010 calibrated value, while all other parameters remain at their baseline levels.

Table 18: Changes in Dynamism due to Less efficient General R&D (in %)

	Implied Values	Contribution
Entry rate	9.06%	91%
Leadership turnover, high-type	13%	86%
Leadership turnover, low-type	12%	65%

Note: Data on entry rate are from BDS. Data on leadership turnover rate are calculated from the firm sample constructed by linking USPTO, PatentsView and Compustat. The second column reports the model-implied values based on the lower efficiency of general R&D investment in post-2010. The third column shows its contribution to the change in business dynamism.

6 Conclusion

This paper studies the role of knowledge generality in shaping market structure, knowledge diffusion, and long-run growth. By distinguishing between general and firm-specific innovations, the model yields two predictions: (1) leaders often limit general knowledge production to preserve their market position, while followers rely on it to absorb external knowledge and narrow the technological gap; and (2) the gap in innovation generality between the two firms varies non-monotonically with the level of market concentration. Using firm-level data, I provide empirical evidence consistent with these patterns. A quasi-natural experiment based on variation in state-level enforceability of non-compete agreements validates the model's key assumption of within-industry spillovers. The quantitative analysis further shows that the post-2010 decline in the efficiency of general R&D investment can account for half of the slowdown in U.S. growth and a large portion of the decline in business dynamism. It also points to a distinct policy implication: well-designed support for general R&D—particularly when directed at leading firms—can amplify spillovers both within and across industries, and sustain long-run growth. Future research could build on this framework in several directions. One promising avenue is to explore the micro-level determinants of generality—such as inventor networks, team diversity, or research collaborations.

References

Abel, Andrew B. (2001) "Will Bequests Attenuate the Predicted Meltdown in Stock Prices When Baby Boomers Retire?" *The Review of Economics and Statistics*, 83 (4), 589–595.

——— (2003) "The Effects of a Baby Boom on Stock Prices and Capital Accumulation in the Presence of Social Security," *Econometrica*, 71 (2), 551–578.

Acemoglu, Daron, Ufuk Akcigit, Harun Alp, Nicholas Bloom, and William Kerr (2018) "Innovation, Reallocation and Growth," *American Economic Review*, 108 (11), 3450–3491.

Acemoglu, Daron, Ufuk Akcigit, and Murat Alp Celik (2022) "Radical and Incremental Innovation: The Roles of Firms, Managers, and Innovators," *American Economic Journal: Macroeconomics*, 14 (3), 199–249.

Acemoglu, Daron and Dan Cao (2015) "Innovation by entrants and incumbents," *Journal of Economic Theory*, 157, 255–294.

Adam, Klaus, Dmitry Matveev, and Stefan Nagel (2021) "Do Survey Expectations of Stock Returns Reflect Risk Adjustments?" *Journal of Monetary Economics*, 117, 723–740.

Aghion, Philippe, Antonin Bergeaud, Timo Boppart, Peter J Klenow, and Huiyu Li (2023) "A Theory of Falling Growth and Rising Rents," *The Review of Economic Studies*, 90 (6), 2675–2702.

Aghion, Philippe, Christopher Harris, Peter Howitt, and John Vickers (2001) "Competition, Imitation and Growth with Step-by-Step Innovation," *The Review of Economic Studies*, 68 (3), 467–492.

Akcigit, Ufuk and Sina T. Ates (2023) "What Happened to US Business Dynamism?" *Journal of Political Economy*, 131 (8), 2059–2124.

Akcigit, Ufuk, Douglas Hanley, and Nicolas Serrano-Velarde (2020) "Back to Basics: Basic Research Spillovers, Innovation Policy, and Growth," *The Review of Economic Studies*, 88 (1), 1–43.

Akcigit, Ufuk and William Kerr (2018) "Growth Through Heterogeneous Innovations," *Journal of Political Economy*, 126 (4).

Aksoy, Cevat, Jose Barrero, Nicholas Bloom, Steven Davis, Mathias Dolls, and Pablo Zarate (2022) "Working From Home Around the World," Brookings Papers on Economic Activity.

Almeida, Paul and Bruce Kogut (1999) "Localization of Knowledge and the Mobility of Engineers in Regional Networks," *Management Science*, 45 (7), 905–917.

Altug, Sumru, Fabrice Collard, Cem Çakmaklı, Sujoy Mukerji, and Han Özsöylev (2020) "Ambiguous Business Cycles: A Quantitative Assessment," *Review of Economic Dynamics*, 38, 220–237.

Ameriks, John and Stephen P. Zeldes (2004) "How Do Household Portfolio Shares Vary With Age?" *mimeo, Columbia University.*

Anderson, Evan W., Lars Peter Hansen, and Thomas J. Sargent (2003) "A Quartet of Semigroups for Model Specification, Robustness, Prices of Risk, and Model Detection," *Journal of the European Economic Association*, 1 (1), 68–123.

Angeletos, George-Marios (2007) "Uninsured idiosyncratic investment risk and aggregate saving," *Review of Economic Dynamics*, 10 (1), 1–30.

Angelici, Marta and Paola Profeta (2023) "Smart Working: Flexibility without Constraints," *Management Science*, 70 (3).

Antoniou, Constantinos, Richard D.F. Harris, and Ruogu Zhang (2015) "Ambiguity Aversion and Stock Market Participation: An Empirical Analysis," *Journal of Banking & Finance*, 58, 57–70.

Argente, David, Salome Baslandze, Douglas Hanley, and Sara Moreira (2023) "Patents to Products: Product Innovation and Firm Dynamics," working paper.

Armantier, Olivier, Luis Armona, Giacomo De Giorgi, and Wilbert van der Klaauw (2016) "Which Households Have Negative Wealth?" *Liberty Street Economics, Federal Reserve Bank of New York.*

Armantier, Olivier, Giorgio Topa, Wilbert van der Klaauw, and Basit Zafar (2017) "An Overview of the Survey of Consumer Expectations," *Federal Reserve Bank of New York Economic Policy Review*, 23 (2).

Arora, Ashish, Sharon Belenzon, and Lia Sheer (2021) "Knowledge Spillovers and Corporate Investment in Scientific Research," *American Economic Review*, 111 (3), 871–98.

Arrow, Kenneth J. (1987) "The Demand for Information and the Distribution of Income," *Probability in the Engineering and Informational Sciences*, 1 (1), 3–13.

Auclert, Adrien, Hannes Malmberg, Frederic Martenet, and Matthew Rognlie (2021) "Demographics, Wealth, and Global Imbalances in the Twenty-First Century," *National Bureau of Economic Research Working Paper Series*, 29161.

Autor, David, Arindrajit Dube, and Annie McGrew (2024) "The Unexpected Compression: Competition at Work in the Low Wage Labor Market," NBER Working Paper 31010.

Babina, Tania, Alex Xi He, Sabrina T Howell, Elisabeth Ruth Perlman, and Joseph Staudt (2023) "Cutting the Innovation Engine: How Federal Funding Shocks Affect University Patenting, Entrepreneurship, and Publications," *The Quarterly Journal of Economics*, 138 (2), 895–954.

Bagga, Sadhika, Lukas Mann, Ayşegül Şahin, and Giovanni L. Violante (2024) "Job Amenity Shocks and Labor Reallocation," working paper.

Balter, Anne, Pascal J. Maenhout, and Hao Xing (2022) "Model Ambiguity versus Model Misspecification in Dynamic Portfolio Choice and Asset Pricing," *mimeo, Tilburg University.*

Barrero, Jose, Nicholas Bloom, and Steven Davis (2021) "Why Working From Home Will Stick," NBER Working Paper 28731.

Barrero, José M., Nick Bloom, and Steven J. Davis (2023) "The Evolution of Work from Home," *Journal of Economic Perspectives*, 37 (4), 1–28.

Barrero, José M., Nick Bloom, Steven J. Davis, Brent Meyer, and Emil Mihaylov (2022) "The Shift to Remote Work Lessens Wage-Growth Pressures," NBER Working Paper 30197.

Basu, Susanto and John G. Fernald (1997) "Returns to Scale in U.S. Production: Estimates and Implications," *Journal of Political Economy*, 105 (2), 249–283.

Battiston, Diego, Jordi Blanes i Vidal, and Tom Kirchmaier (2021) "Face-to-Face Communication in Organizations," *Review of Economic Studies*, 88 (2), 574–609.

Benhabib, Jess, Jesse Perla, and Christopher Tonetti (2021) "Reconciling Models of Diffusion and Innovation: A Theory of the Productivity Distribution and Technology Frontier," *Econometrica*, 89 (5), 2261–2301.

Bessen, James E., Erich Denk, Joowon Kim, and Cesare Righi (2020) "Declining Industrial Disruption," working paper, Boston Univ. School of Law, Law and Economics Research Paper 20-28.

Bhutta, Neil, Jesse Bricker, Andrew C. Chang et al. (2020) "Changes in U.S. Family Finances from 2016 to 2019: Evidence from the Survey of Consumer Finances," *Federal Reserve Bulletin*, 106 (5), 1–42.

Bidder, R. M. and M. E. Smith (2012) "Robust Animal Spirits," *Journal of Monetary Economics*, 59 (8), 738–750.

Bishara, Norman D. (2011) "Fifty Ways to Leave Your Employer: Relative Enforcement of Covenants Not to Compete, Trends, and Implications for Employee Mobility Policy," *University of Pennsylvania Journal of Business Law*, 13.

Bloom, Nicholas, Charles I. Jones, John Van Reenen, and Michael Webb (2020) "Are Ideas Getting Harder to Find?" *American Economic Review*, 110 (4), 1104–44.

Bloom, Nicholas, James Liang, John Roberts, and Zhichun Jenny Ying (2015) "Does Working from Home Work? Evidence from a Chinese Experiment," *Quarterly Journal of Economics*, 130 (1), 165–218.

Bloom, Nicholas, Mark Schankerman, and John Van Reenen (2013) "Identifying Technology Spillovers and Product Market Rivalry," *Econometrica*, 81 (4), 1347–1393.

Bloom, Nick, Rachel Griffith, and John Van Reenen (2002) "Do R&D tax credits work? Evidence from a panel of countries 1979–1997," *Journal of Public Economics*, 85, 1–31.

Bloom, Nick, Ruobing Han, and James Liang (2024) "Hybrid working from home improves retention without damaging performance," *Nature*, 630 (8018), 920–925.

Boedo, Hernan and Toshihiko Mukoyama (2012) "Evaluating the Effects of Entry Regulations and Firing Costs on International Income Differences," *Journal of Economic Growth*, 17, 143–170.

Briggs, Joseph, David Cesarini, Erik Lindqvist, and Robert Östling (2020) "Windfall Gains and Stock Market Participation," *Journal of Financial Economics*, 139, 57–83.

Cagetti, Marco, Lars Peter Hansen, Thomas Sargent, and Noah Williams (2002) "Robustness and Pricing with Uncertain Growth," *The Review of Financial Studies*, 15 (2), 363–404.

Caggese, Andrea (2019) "Financing Constraints, Radical versus Incremental Innovation, and Aggregate Productivity," *American Economic Journal: Macroeconomics*, 11 (2), 275–309.

Caicedo, Santiago and Jeremy Pearce (2024) "Need for Speed: Quality of Innovations and the Allocation of Inventors," working paper.

Caliendo, Frank N., Aspen Gorry, and Sita Slavov (2020) "Survival Ambiguity and Welfare," *Journal of Economic Behavior & Organization*, 170, 20–42.

Campanale, Claudio (2011) "Learning, Ambiguity and Life-Cycle Portfolio Allocation," *Review of Economic Dynamics*, 14 (2), 339–367.

Campbell, John Y. (1993) "Intertemporal Asset Pricing without Consumption Data," *The American Economic Review*, 83 (3), 487–512.

Campbell, John Y. and Luis M. Viceira (2002) *Strategic Asset Allocation: Portfolio Choice for Long-Term Investors*: Oxford University Press.

Carvalho, Carlos, Andrea Ferrero, and Fernanda Nechio (2016) "Demographics and Real Interest Rates: Inspecting the Mechanism," *European Economic Review*, 88, 208–226.

Casal, Lucía (2025) "Lock-In and Productive Innovations: Implications for Firm-to-Firm Innovation Pass-Through," working paper.

Catherine, Sylvain (2022) "Countercyclical Labor Income Risk and Portfolio Choices over the Life Cycle," *The Review of Financial Studies*, 35 (9), 4016–4054.

Chang, Yongsung, Jay H. Hong, and Marios Karabarbounis (2018) "Labor Market Uncertainty and Portfolio Choice Puzzles," *American Economic Journal: Macroeconomics*, 10 (2), 222–262.

Che, Yeon-Koo, Elisabetta Iossa, and Patrick Rey (2021) "Prizes versus Contracts as Incentives for Innovation," *The Review of Economic Studies*, 88 (5), 2149–2178.

Chen, Hui, Nengjiu Ju, and Jianjun Miao (2014) "Dynamic Asset Allocation with Ambiguous Return Predictability," *Review of Economic Dynamics*, 17 (4), 799–823.

Choudhury, Prithwiraj, Cirrus Foroughi, and Barbara Zepp Larson (2021) "Work-from-Anywhere: The Productivity Effects of Geographic Flexibility," *Startegic Management Journal*, 42 (4), 655–683.

Clementi, Gian Luca and Berardino Palazzo (2016) "Entry, Exit, Firm Dynamics, and Aggregate Fluctuations," *American Economic Journal: Macroeconomics*, 8 (3), 1–41.

Cocco, João F., Francisco J. Gomes, and Pascal J. Maenhout (2005) "Consumption and Portfolio Choice over the Life Cycle," *The Review of Financial Studies*, 18 (2), 491–533.

Cohen, Wesley and Daniel Levinthal (1990) "Absorptive Capacity: A New Perspective on Learning and Innovation," *Administrative Science Quarterly*, 35, 128–152.

Collard, Fabrice, Sujoy Mukerji, Kevin Sheppard, and Jean-Marc Tallon (2018) "Ambiguity and the Historical Equity Premium," *Quantitative Economics*, 9 (2), 945–993.

Cooper, Russell W. and John C. Haltiwanger (2006) "On the Nature of Capital Adjustment Costs," *The Review of Economic Studies*, 73 (3), 611–633.

Corgnet, Brice, Roberto Hernán-González, and Praveen Kujal (2020) "On Booms That Never Bust: Ambiguity in Experimental Asset Markets with Bubbles," *Journal of Economic Dynamics and Control*, 110, 103754.

Crump, Richard K., Stefano Eusepi, Andrea Tambalotti, and Giorgio Topa (2022) "Subjective Intertemporal Substitution," *Journal of Monetary Economics*, 126 (March), 118–133.

Cunningham, Colleen, Florian Ederer, and Song Ma (2021) "Killer Acquisitions," *Journal of Political Economy*, 129 (3), 649–702.

Davis, Morris, Andra Ghent, and Jesse Gregory (2024) "The Work-From-Home Technology Boom and its Consequences," *The Review of Economic Studies*, rdad114.

De Loecker, Jan, Jan Eeckhout, and Simon Mongey (2021) "Quantifying Market Power and Business Dynamism in the Macroeconomy," working paper, National Bureau of Economic Research.

De Ridder, Maarten (2024) "Market Power and Innovation in the Intangible Economy," *American Economic Review*, 114 (1), 199–251.

Decker, Ryan A. and John Haltiwanger (2024) "Surging business formation in the pandemic: Causes and consequences?" *Brookings Papers on Economic Activity*.

DellaVigna, Stefano and Joshua M. Pollet (2007) "Demographics and Industry Returns," *American Economic Review*, 97 (5), 1667–1702.

Dimmock, Stephen G., Roy Kouwenberg, Olivia S. Mitchell, and Kim Peijnenburg (2016) "Ambiguity Aversion and Household Portfolio Choice Puzzles: Empirical Evidence," *Journal of Financial Economics*, 119 (3), 559–577.

Dingel, Jonathan and Brent Neiman (2020) "How Many Jobs Can Be Done from Home?" *Journal of Public Economics*, 189.

Dyévre, Arnaud (2024) "Public R&D Spillovers and Productivity Growth," working paper.

Eggertsson, Gauti B., Neil R. Mehrotra, and Jacob A. Robbins (2019) "A Model of Secular Stagnation: Theory and Quantitative Evaluation," *American Economic Journal: Macroeconomics*, 11 (1), 1–48.

Emanuel, Natalia and Emma Harrington (2023) "Working Remotely? Selection, Treatment, and the Market for Remote Work," Federal Reserve Bank of New York Staff Report 1061.

Epstein, Larry G. and Martin Schneider (2010) "Ambiguity and Asset Markets," *Annual Review of Financial Economics*, 2 (1), 315–346.

Fernald, John G. (2015) "Productivity and Potential Output before, during, and after the Great Recession," *NBER Macroeconomics Annual*, 29 (1), 1–51.

Fernández-Villaverde, Jesús, Yang Yu, and Francesco Zanetti (2025) "Defensive Hiring and Creative Destruction," CESifo Working Paper 11753.

Ferriere, Axelle and Anastasios G. Karantounias (2019) "Fiscal Austerity in Ambiguous Times," *American Economic Journal: Macroeconomics*, 11 (1), 89–131.

Flynn, Joel P., Lawrence D. W. Schmidt, and Alexis Akira Toda (2023) "Robust Comparative Statics for the Elasticity of Intertemporal Substitution," *Theoretical Economics*, 18 (1), 231–265.

Foltyn, Richard (2025) "Experience-Based Learning, Stock Market Participation and Portfolio Choice," *mimeo, Norwegian School of Economics*.

Foltyn, Richard and Jonna Olsson (2024) "Subjective Life Expectancies, Time Preference Heterogeneity, and Wealth Inequality," *Quantitative Economics*, 15, 699–736.

Fossum, Eric R. (2020) "The Invention of CMOS Image Sensors: A Camera in Every Pocket," in *2020 Pan Pacific Microelectronics Symposium (Pan Pacific)*, 1–6.

Foster, Lucia, John Haltiwanger, and Chad Syverson (2008) "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" *American Economic Review*, 98 (1), 394–425.

Gagnon, Etienne, Benjamin K. Johannsen, and David López-Salido (2021) "Understanding the New Normal: The Role of Demographics," *IMF Economic Review*, 69 (2), 357–390.

Garcia-Macia, Daniel, Chang-Tai Hsieh, and Peter J. Klenow (2019) "How Destructive Is Innovation?" *Econometrica*, 87 (5), 1507–1541.

Garlappi, Lorenzo, Raman Uppal, and Tan Wang (2007) "Portfolio Selection with Parameter and Model Uncertainty: A Multi-Prior Approach," *The Review of Financial Studies*, 20 (1), 41–81.

Geanakoplos, John, Michael Magill, and Martine Quinzii (2004) "Demography and the Long-Run Predictability of the Stock Market," *Brookings Papers on Economic Activity*, 2004, 241–307.

Gibbs, Michael, Friederike Mengel, and Christoph Siemroth (2023) "Work from Home and Productivity: Evidence from Personnel and Analytics Data on Information Technology Professionals," *Journal of Political Economy Microeconomics*, 1 (1).

Giglio, Stefano, Matteo Maggiori, Johannes Stroebel, and Stephen Utkus (2021) "Five Facts about Beliefs and Portfolios," *American Economic Review*, 111 (5), 1481–1522.

Gilboa, Itzhak and David Schmeidler (1989) "Maxmin Expected Utility with Non-Unique Prior," *Journal of Mathematical Economics*, 18 (2), 141–153.

Giroud, Xavier, Ernest Liu, and Holger Mueller (2024) "Innovation Spillovers across U.S. Tech Clusters," working paper.

Gomes, Francisco (2020) "Portfolio Choice Over the Life Cycle: A Survey," *Annual Review of Financial Economics*, 12 (1), 277–304.

Gomes, Francisco, Michael Haliassos, and Tarun Ramadorai (2021) "Household Finance," *Journal of Economic Literature*, 59 (3), 919–1000.

Gomes, Francisco and Alexander Michaelides (2005) "Optimal Life-Cycle Asset Allocation: Understanding the Empirical Evidence," *The Journal of Finance*, 60 (2), 869–904.

Gomes, Francisco and Oksana Smirnova (2021) "Stock Market Participation and Portfolio Shares Over the Life-Cycle," *mimeo, London Business School.*

Goodhart, Charles and Manoj Pradhan (2020) *The Great Demographic Reversal: Ageing Societies, Waning Inequality, and an Inflation Revival*: Palgrave Macmillan, 1st edition.

Gorodnichenko, Yuriy and Michael Weber (2016) "Are Sticky Prices Costly? Evidence from the Stock Market," *American Economic Review*, 106 (1).

Graff Zivin, Joshua and Elizabeth Lyons (2021) "The Effects of Prize Structures on Innovative Performance," *AEA Papers and Proceedings*, 111, 577–81.

Grevenbrock, Nils, Max Groneck, Alexander Ludwig, and Alexander Zimper (2021) "Cognition, Optimism, and the Formation of Age-Dependent Survival Beliefs," *International Economic Review*, 62 (2), 887–918.

Hall, Bronwyn H, Adam B Jaffe, and Manuel Trajtenberg (2001) "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools," Working Paper 8498, National Bureau of Economic Research.

Haltiwanger, John, Ron Jarmin, Robert Kulick, and Javier Miranda (2016) "High Growth Young Firms: Contribution to Job, Output and Productivity Growth," mimeo.

Hamermesh, Daniel S., Harley Frazis, and Jay Stewart (2005) "Data Watch: The American Time Use Survey," *Journal of Economic Perspectives*, 19 (1), 221–232.

Hansen, Lars Peter and Thomas J Sargent (2008) *Robustness*: Princeton University Press.

Hansen, Lars Peter, Thomas J. Sargent, and Thomas D. Tallarini, Jr. (1999) "Robust Permanent Income and Pricing," *The Review of Economic Studies*, 66 (4), 873–907.

Hansen, Stephen, Peter Lamber, Nicholas Bloom, Steven Davis, Raffaella Sadun, and Bledi Taska (2023) "Remote Work Across Jobs, Companies, and Space," NBER Working Paper 31007.

Havránek, Tomáš (2015) "Measuring Intertemporal Substitution: The Importance of Method Choices and Selective Reporting," *Journal of the European Economic Association*, 13 (6), 1180–1204.

He, Haoran, David Neumark, and Qian Weng (2021) "Do Workers Value Flexible Jobs? A Field Experiment," *Journal of Labor Economics*, 39 (3).

Hopenhayn, Hugo and Richard Rogerson (1993) "Job Turnover and Policy Evaluation: A General Equilibrium Analysis," *Journal of Political Economy*, 101 (5), 915–938.

Ignaszak, Marek and Petr Sedláček (2023) "Customer Acquisition, Business Dynamism and Aggregate Growth," mimeo.

Ilut, Cosmin L. and Martin Schneider (2023) "Modeling Uncertainty as Ambiguity: A Review," in Bachmann, Ruediger, Giorgio Topa, and Wilbert Van Der Klaauw eds. *Handbook of Economic Expectations*: Elsevier.

Jaffe, Adam B., Manuel Trajtenberg, and Michael S. Fogarty (2000) "Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors," *The American Economic Review*, 90 (2), 215–218.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson (1993) "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," *The Quarterly Journal of Economics*, 108 (3).

Klette, Tor Jakob and Samuel Kortum (2004) "Innovating Firms and Aggregate Innovation," *Journal of Political Economy*, 112 (5), 986–1018.

Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji (2005) "A Smooth Model of Decision Making under Ambiguity," *Econometrica*, 73 (6), 1849–1892.

——— (2009) "Recursive Smooth Ambiguity Preferences," *Journal of Economic Theory*, 144 (3), 930–976.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman (2017) "Technological Innovation, Resource Allocation, and Growth," *The Quarterly Journal of Economics*, 132 (2), 665–712.

Kopecky, Joseph and Alan M. Taylor (2022) "The Savings Glut of the Old: Population Aging, the Risk Premium, and the Murder-Suicide of the Rentier," *National Bureau of Economic Research Working Paper Series*.

Kovalchik, Stephanie, Colin F. Camerer, David M. Grether, Charles R. Plott, and John M. Allman (2005) "Aging and Decision Making: A Comparison between Neurologically Healthy Elderly and Young Individuals," *Journal of Economic Behavior & Organization*, 58 (1), 79–94.

Krause, Simon, Andreas Trumpp, Tobias Dichtl, Susanne Kiese, and Alexander Rutsch (2024) "The Impact of Working from Home on the German Office Real Estate Market," EconPol Forum, vol. 25, CES-Ifo.

Kroen, Thomas, Ernest Liu, Atif R Mian, and Amir Sufi (2021) "Falling Rates and Rising Superstars," Working Paper 29368, National Bureau of Economic Research.

Kuvshinov, Dmitry and Kaspar Zimmermann (2020) "The Expected Return on Risky Assets: International Long-run Evidence," *CEPR Discussion Papers*, 15610.

Lei, Xiaowen (2019) "Information and Inequality," *Journal of Economic Theory*, 184, 104937.

León-Ledesma, Miguel and Mathan Satchi (2019) "Appropriate Technology and Balanced Growth," *Review of Economic Studies*, 86, 807–835.

Li, Huiyu, Julien Sauvagnat, and Tom Schmitz (2026) "The Work-from-Home Wage Premium," *CEPR Discussion Paper No. 20996.*

Lin, Yiling, Carl Benedikt Frey, and Lingfei Wu (2023) "Remote Collaboration Fuses Fewer Breakthrough Ideas," *Nature*, 623, 987–991.

Liu, Ernest and Song Ma (2021) "Innovation Networks and RD Allocation," Working Paper 29607, National Bureau of Economic Research.

Liu, Ernest, Atif Mian, and Amir Sufi (2022) "Low Interest Rates, Market Power, and Productivity Growth," *Econometrica*, 90 (1), 193–221.

Liu, Jingnan (2023) "Worker Mobility, Knowledge Diffusion, and Non-Compete Contracts," working paper.

Liu, Sitian and Yichen Su (2024) "The Effect of Working from Home on the Agglomeration Economies of Cities: Evidence from Advertised Wages," mimeo.

Ma, Song, Wenyu Wang, and Yufeng Wu (2025) "Steering Labor Mobility Through Innovation," working paper, Fisher College of Business.

Macaulay, Alistair (2021) "The Attention Trap: Rational Inattention, Inequality, and Fiscal Policy," *European Economic Review*, 135, 103716.

Maenhout, Pascal J. (2004) "Robust Portfolio Rules and Asset Pricing," *The Review of Financial Studies*, 17 (4), 951–983.

Malmendier, Ulrike, Demian Pouzo, and Victoria Vanasco (2020) "Investor Experiences and Financial Market Dynamics," *Journal of Financial Economics*, 136 (3), 597–622.

Mankiw, N. Gregory and David N. Weil (1989) "The Baby Boom, the Baby Bust, and the Housing Market," *Regional Science and Urban Economics*, 19 (2), 235–258.

Marx, Matt (2022) "Employee Non-compete Agreements, Gender, and Entrepreneurship," *Organization Science*, 33 (5), 1756–1772.

Marx, Matt, Deborah Strumsky, and Lee Fleming (2009) "Mobility, Skills, and the Michigan Non-Compete Experiment," *Management Science*, 55 (6), 875–889.

Mas, Alexandre and Amanda Pallais (2017) "Valuing Alternative Work Arrangements," *American Economic Review*, 107 (12), 3722–59.

Merton, Robert C. (1969) "Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case," *The Review of Economics and Statistics*, 51 (3), 247–257.

Mian, Atif, Ludwig Straub, and Amir Sufi (2021) "What Explains the Decline in R*? Rising Income Inequality Versus Demographic Shifts," *Jackson Hole Conference Proceedings*.

Michelacci, Claudio and Luigi Paciello (2020) "Ambiguous Policy Announcements," *The Review of Economic Studies*, 87 (5), 2356–2398.

——— (2023) "Ambiguity Aversion and Heterogeneity in Households' Beliefs," *American Economic Journal: Macroeconomics*.

Monte, Ferdinando, Charly Porcher, and Esteban Rossi-Hansberg (2023) "Remote Work and City Structure," Working Paper 31494, National Bureau of Economic Research.

Natalia, Monteiro, Straume Odd Rune, and Valente Marieta (2019) "Does Remote Work Improve or Impair Firm Labour Productivity? Longitudinal Evidence from Portugal," CESifo Working Paper 7991.

Olmstead-Rumsey, Jane (2025) "Market Concentration and the Productivity Slowdown," working paper.

Peijnenburg, Kim (2018) "Life-Cycle Asset Allocation with Ambiguity Aversion and Learning," *Journal of Financial and Quantitative Analysis*, 53 (5), 1963–1994.

Perla, Jesse and Christopher Tonetti (2014) "Equilibrium Imitation and Growth," *Journal of Political Economy*, 122 (1), 52–76.

Peters, Michael (2020) "Heterogeneous Markups, Growth, and Endogenous Misallocation," *Econometrica*, 88 (5), 2037–2073.

Poschke, Markus (2010) "The Regulation of Entry and Aggregate Productivity," *The Economic Journal*, 120, 1175–1200.

Poterba, James M. (2001) "Demographic Structure and Asset Returns," *The Review of Economics and Statistics*, 83 (4), 565–584.

Reinmuth, Kate and Emma Rockall (2025) "Innovation through Inventor Mobility: Evidence from Non-Compete Agreements," working paper.

Richard, Morgane (2024) "The Spatial and Distributive Implications of Working-from-Home: A General Equilibrium Model," mimeo.

Rudebusch, Glenn D and Eric T Swanson (2012) "The Bond Premium in a DSGE Model with Long-Run Real and Nominal Risks," *American Economic Journal: Macroeconomics*, 4 (1), 105–143.

Samila, Sampsa and Olav Sorenson (2011) "Noncompete Covenants: Incentives to Innovate or Impediments to Growth," *Management Science*, 57 (3), 425–438.

Sedláček, Petr (2020) "Lost Generations of Firms and Aggregate Labor Market Dynamics," *Journal of Monetary Economics*, 111, 16–31.

Sedláček, Petr and Vincent Sterk (2017) "The Growth Potential of Startups over the Business Cycle," *American Economic Review*, 107 (10), 3182–3210.

Singh, Jasjit and Ajay Agrawal (2011) "Recruiting for Ideas: How Firms Exploit the Prior Inventions of New Hires," *Management Science*, 57 (1), 129–150.

Stoyanov, Andrey and Nikolay Zubanov (2012) "Productivity Spillovers across Firms through Worker Mobility," *American Economic Journal: Applied Economics*, 4 (2), 168–98.

Straub, Ludwig (2019) "Consumption, Savings, and the Distribution of Permanent Income," *mimeo, Harvard University.*

Trajtenberg, Manuel, Rebecca Henderson, and Adam Jaffe (1997) "University Versus Corporate Patents: A Window On The Basicness Of Invention," *Economics of Innovation and New Technology*, 5 (1), 19–50.

Vij, Akshay, Flavio Souza, Helen Barrie, V Anilan, Sergio Sarmiento, and Lynette Washington (2023) "Employee Preferences for Working From Home in Australia," *Journal of Economic Behavior and Organization*, 214.

Yang, Longqi, David Holtz, Sonia Jaffee, Siddharth Suri, Shilpi Sinha, Jeffrey Weston, and Connor Joyce et al. (2022) "The Effects of Remote Work on Collaboration among Information Workers," *Nature Human Behavior*, 6, 43–54.

Yogo, Motohiro (2016) "Portfolio Choice in Retirement: Health Risk and the Demand for Annuities, Housing, and Risky Assets," *Journal of Monetary Economics*, 80, 17–34.

Zarate, Pablo, Mathias Dolls, Steven Davis, Nicholas Bloom, Jose Maria Barrero, and Cevat Giray Aksoy (2024) "Why Does Working From Home Vary Across Countries and People?," NBER Working Paper 32374.

Appendix: Chapter 1

A Proofs

Proposition 1: Since $\phi_J = 0$ and there are no bequests, there is no incentive to save at age J. As we also impose that $w_{J+1,T+1} \geq 0$ (10), all agents at age J choose to consume all of their wealth. With no continuation value, the value function of these agents is:

$$V_J(w_{j,t}) = \frac{w_{j,t}^{1-\gamma}}{1-\gamma} \tag{115}$$

For all $j < J$, we now conjecture that the value function takes the same functional form:

$$V_j(w_{j,t}) = A_j \frac{w_{j,t}^{1-\gamma}}{1-\gamma} \tag{116}$$

for some age-dependent constant A_j. The expectation of the next-period value function becomes:

$$E_{j,t}[V_{j+1}(w_{j+1,t+1})] = E_{j,t}[A_{j+1}\frac{w_{j+1}^{1-\gamma}}{1-\gamma}] \tag{117}$$

$$\approx \frac{A_{j+1}}{1-\gamma}(w_{j,t}-c_{j,t})^{1-\gamma} E_{j,t}\left\{\exp\left[r^f + \alpha_{j,t}(r_{t+1}-r^f) + \frac{1}{2}\alpha_{j,t}(1-\alpha_{j,t})\sigma^2\right]\right\}^{1-\gamma} \tag{118}$$

where equation (118) uses the log-linear approximation to the one-period excess return around zero in equation (12) and Campbell (1993). Since r_{t+1} is normally distributed, we can further write:

$$E_{j,t}[V_{j+1}(w_{j+1,t+1})] = \frac{A_{j+1}}{1-\gamma}(w_{j,t}-c_{j,t})^{1-\gamma} \exp\left\{(1-\gamma)\left[r^f + \alpha_{j,t}(\widetilde{\mu}-r^f) + \frac{1}{2}\alpha_{j,t}(1-\alpha_{j,t})\sigma^2\right] + \frac{1}{2}(1-\gamma)^2\alpha_{j,t}^2\sigma^2\right\} \tag{119}$$

Using this in equation (8) and taking first order conditions with respect to $\alpha_{j,t}, c_{j,t}$ gives equations (17) and (18), with b_{j+1} as in equation (16) (noting that $\mu \equiv \tilde{\mu} + \sigma^2/2$).

Finally, we verify the conjecture in equation (116) for $j = 0, ..., J-1$. Substituting equation (119) into the value function (8) we have:

$$A_j \frac{w_{j,t}^{1-\gamma}}{1-\gamma} = \frac{(c_{j,t}^*)^{1-\gamma}}{1-\gamma}$$
$$+ \beta\phi_j \frac{A_{j+1}}{1-\gamma}(w_{j,t} - c_{j,t}^*)^{1-\gamma} \left\{ 1 + (1-\gamma)[r^f + (\mu - r^f)\alpha_{j,t}^*] - \frac{1}{2}\gamma(1-\gamma)\sigma^2(\alpha^*)^2 \right\} \quad (120)$$

where $\alpha^*, c_{j,t}^*$ denote the optimal risky share and consumption from equations (17) and (18) respectively. Since α^* is a constant, and $c_{j,t}^*$ is proportional to $w_{j,t}$, this verifies the conjectured functional form for $V_j(w_{j,t})$ (116). Matching coefficients, we obtain equation (15).

Lemma 1: First conjecture that the value function takes the form:

$$V_j^\theta(w_{j,t}) = A_j \frac{w_{j,t}^{1-\gamma}}{1-\gamma} + \theta B_j \frac{w_{j,t}^{2(1-\gamma)}}{2(1-\gamma)} + O(\theta^2) \quad (121)$$

Taking the same log-linear approximation approach as in (12) to the distorted returns in (19), we can write

$$\begin{aligned} E_{j,t}[V_{j+1}^\theta(w_{j+1,t+1})] \approx\ & \frac{A_{j+1}}{1-\gamma}(w_{j,t} - c_{j,t})^{1-\gamma}\{1 + (1-\gamma)[r^f + (\widetilde{\mu} - r^f + \sigma\nu_{j,t})\alpha_{j,t}] \\ & + \frac{1}{2}(1-\gamma)\sigma^2(\alpha_{j,t} - \gamma\alpha_{j,t}^2)\} \\ & + \theta\frac{B_{j+1}}{2(1-\gamma)}(w_{j,t} - c_{j,t})^{2(1-\gamma)}\{1 + 2(1-\gamma)[r^f + (\widetilde{\mu} - r^f + \sigma\nu_{j,t})\alpha_{j,t}] \\ & + (1-\gamma)\sigma^2(\alpha_{j,t} + \alpha_{j,t}^2 - 2\gamma\alpha_{j,t}^2)\} \\ =\ & E[V_{j+1}(w_{j+1,t+1}^*)] + \frac{A_{j+1}}{1-\gamma}(w_{j,t} - c_{j,t})^{1-\gamma}(1-\gamma)\sigma\alpha_{j,t}\nu_{j,t} \\ & + \theta\frac{B_{j+1}}{2(1-\gamma)}(w_{j,t} - c_{j,t})^{2(1-\gamma)}2(1-\gamma)\sigma\alpha_{j,t}\nu_{j,t} \end{aligned} \quad (122)$$

where in the first approximation we drop the term with θ^2 and higher orders, and use the approximation:

$$R_{t+1} + \sigma_1\nu_{j,t} \approx \exp(r_{t+1} + \sigma\nu_{j,t}) \quad (123)$$

In equation (123), σ_1 is the standard deviation of R_{t+1}, which equals to $[\exp(\sigma^2) - 1]^{\frac{1}{2}}[\exp(2\widetilde{\mu} + \sigma^2)]^{\frac{1}{2}}$. Note that $\theta\nu_{j,t} << \nu_{j,t}$, we can further approximate $E_{j,t}[V_{j+1}^\theta(w_{j+1,t+1})]$ as:

$$E_{j,t}[V_{j+1}^\theta(w_{j+1,t+1})] \approx E_{j,t}[V_{j+1}(w_{j+1,t+1}^*)] + \frac{A_{j+1}}{1-\gamma}(w_{j,t} - c_{j,t})^{1-\gamma}(1-\gamma)\sigma\alpha_{j,t}\nu_{j,t} \quad (124)$$

Proposition 2: First, take equation (124) (identical to equation (22) in Lemma 1), and substitute out for $\nu_{j,t}$ using equation (24), to obtain:

$$E_{j,t}[V_{j+1}^{\theta}(w_{j+1,t+1})] \approx E_{j,t}[V_{j+1}(w_{j+1,t+1}^{*})] - \theta A_{j+1}^{2}(w_{j,t} - c_{j,t})^{2-2\gamma}\sigma^{2}\alpha_{j,t}^{2} \tag{125}$$

Using this to substitute out for the continuation value in the Bellman equation (equation (25)) yields:

$$\begin{aligned} V_j^{\theta}(w_{j,t}) &= \max_{c_{j,t},\alpha_{j,t}} \left\{ \frac{c_{j,t}^{1-\gamma}}{1-\gamma} + \beta\phi_j \left[\frac{1}{2}\theta A_{j+1}^{2}(w_{j,t} - c_{j,t})^{2-2\gamma}\sigma^{2}\alpha_{j,t}^{2} + E_{j,t}[V_{j+1}^{\theta}(w_{j+1,t+1})] \right] \right\} \\ &\approx \max_{c_{j,t},\alpha_{j,t}} \left\{ \frac{c_{j,t}^{1-\gamma}}{1-\gamma} + \beta\phi_j \left[-\frac{1}{2}\theta A_{j+1}^{2}(w_{j,t} - c_{j,t})^{2-2\gamma}\sigma^{2}\alpha_{j,t}^{2} + E_{j,t}[V_{j+1}^{\theta}(w_{j+1,t+1}^{*})] \right] \right\} \end{aligned} \tag{126}$$

Continue with the guess of the value function's form:

$$V_j^{\theta}(w_{j,t}) = A_j \frac{w_{j,t}^{1-\gamma}}{1-\gamma} + \theta B_j \frac{w_{j,t}^{2(1-\gamma)}}{2(1-\gamma)} + O(\theta^2) \tag{127}$$

$$\approx V_j^{0}(w_{j,t}) + \theta V_j^{1}(w_{j,t}) \tag{128}$$

Note that A_j is the same as in the benchmark model. When there is no ambiguity aversion (i.e. $\theta = 0$), the value function degenerates into that of the benchmark model.

It is intuitive to conjecture that the optimal portfolio choice and consumption take the following forms:

$$\alpha_{j,t} = \alpha^{*} + \theta\alpha'_{j,t} \tag{129}$$

$$c_{j,t} = c_{j,t}^{*} + \theta c'_{j,t} \tag{130}$$

where $\alpha^{*}, c_{j,t}^{*}$ are the solutions without ambiguity defined in Proposition 1.

Then we can approximate the value function by dropping the term with θ^2 and higher orders:

$$\begin{aligned} V_j^{\theta}(w_{j,t}) &\approx \max_{c,\alpha}\{\frac{c_{j,t}^{1-\gamma}}{1-\gamma} + \beta\phi_j[-\frac{1}{2}\theta A_{j+1}^{2}(w_{j,t} - c_{j,t})^{2-2\gamma}\sigma^{2}\alpha_{j,t}^{2} + E_{j,t}[V_{j+1}^{\theta}(w_{j+1,t+1}^{*})]]\} \\ &\approx \max_{c,\alpha}\{\frac{c_{j,t}^{1-\gamma}}{1-\gamma} + \beta\phi_j[-\frac{1}{2}\theta A_{j+1}^{2}(w_{j,t} - c_{j,t})^{2-2\gamma}\sigma^{2}\alpha_{j,t}^{2} + E_{j,t}[V_{j+1}^{0}(w_{j+1,t+1}^{*})] \\ &\quad +\theta E_{j,t}[V_{j+1}^{1}(w_{j+1,t+1}^{*})]]\} \end{aligned} \tag{131}$$

where using log linearization we have:

$$
\begin{aligned}
E_{j,t}[V^0_{j+1}(w^*_{j+1,t+1})] &\approx \frac{A_{j+1}}{1-\gamma}(w_{j,t}-c_{j,t})^{1-\gamma}\{1+(1-\gamma)[r^f+(\widetilde{\mu}-r^f)\alpha_{j,t}] \\
&\quad +\frac{1}{2}(1-\gamma)\sigma^2(\alpha_{j,t}-\gamma\alpha^2_{j,t})\} \\
E_{j,t}[V^1_{j+1}(w^*_{j+1,t+1})] &\approx \frac{B_{j+1}}{2(1-\gamma)}(w_{j,t}-c_{j,t})^{2(1-\gamma)}\{1+2(1-\gamma)[r^f+(\widetilde{\mu}-r^f)\alpha_{j,t}] \\
&\quad +(1-\gamma)\sigma^2(\alpha_{j,t}+\alpha^2_{j,t}-2\gamma\alpha^2_{j,t})\}
\end{aligned}
$$

The FOC w.r.t. $\alpha_{j,t}$ gives:

$$
\begin{aligned}
0 &= -\theta A^2_{j+1}(w_{j,t}-c_{j,t})^{1-\gamma}\sigma^2\alpha_{j,t}+A_{j+1}\{(\widetilde{\mu}-r^f)+\frac{1}{2}\sigma^2-\gamma\sigma^2\alpha_{j,t}\} \\
&\quad +\theta B_{j+1}(w_{j,t}-c_{j,t})^{1-\gamma}\{(\widetilde{\mu}-r^f)+\frac{1}{2}\sigma^2+(1-2\gamma)\sigma^2\alpha_{j,t}\}
\end{aligned}
$$

Substituting $\mu=\widetilde{\mu}+\frac{1}{2}\sigma^2$ into the equation gives:

$$
\begin{aligned}
0 &= -\theta A^2_{j+1}(w_{j,t}-c_{j,t})^{1-\gamma}\sigma^2\alpha_{j,t}+A_{j+1}\{(\mu-r^f)-\gamma\sigma^2\alpha_{j,t}\} \\
&\quad +\theta B_{j+1}(w_{j,t}-c_{j,t})^{1-\gamma}\{(\mu-r^f)+(1-2\gamma)\sigma^2\alpha_{j,t}\}
\end{aligned}
$$

By plugging $\alpha_{j,t}=\alpha^*_{j,t}+\theta\alpha'_{j,t}=\frac{\mu-r}{\gamma\sigma^2}+\theta\alpha'_{j,t}$, we get:

$$
\begin{aligned}
0 &= -\theta A^2_{j+1}(w_{j,t}-c_{j,t})^{1-\gamma}\sigma^2(\alpha^*_{j,t}+\theta\alpha'_{j,t})-\theta A_{j+1}\gamma\sigma^2\alpha'_{j,t} \\
&\quad +\theta B_{j+1}(w_{j,t}-c_{j,t})^{1-\gamma}\{(\mu-r^f)+(1-2\gamma)\sigma^2(\frac{\mu-r^f}{\gamma\sigma^2}+\theta\alpha'_{j,t})\}
\end{aligned}
$$

Further approximation by dropping terms with θ^2 and simplification give:

$$
\begin{aligned}
0 &\approx -A^2_{j+1}(w_{j,t}-c_{j,t})^{1-\gamma}\sigma^2\alpha^*_{j,t}-A_{j+1}\gamma\sigma^2\alpha'_{j,t} \\
&\quad +B_{j+1}(w_{j,t}-c_{j,t})^{1-\gamma}(\mu-r^f)(\frac{1}{\gamma}-1) \\
\alpha'_{j,t} &= -\frac{A^2_{j+1}+B_{j+1}(\gamma-1)}{A_{j+1}\gamma}\alpha^*_{j,t}(w_{j,t}-c^*_{j,t})^{1-\gamma} \\
&= -\frac{A^2_{j+1}+B_{j+1}(\gamma-1)}{A_{j+1}\gamma}\alpha^*_{j,t}(\frac{w_{j,t}}{1+b_{j+1}})^{1-\gamma} \qquad (132)
\end{aligned}
$$

Then we can simplify $E_{j,t}[V^0_{j+1}(w^*_{j+1,t+1})]$ and $E_{j,t}[V^1_{j+1}(w^*_{j+1,t+1})]$:

$$\begin{aligned}
E_{j,t}[V^0_{j+1}(w^*_{j+1,t+1})] &\approx \frac{A_{j+1}}{1-\gamma}(w_{j,t}-c_{j,t})^{1-\gamma}\{1+(1-\gamma)r^f+\frac{1}{2}\gamma(1-\gamma)\sigma^2(\alpha^*_{j,t})^2\} \\
E_{j,t}[V^1_{j+1}(w^*_{j+1,t+1})] &\approx \frac{B_{j+1}}{2(1-\gamma)}(w_{j,t}-c_{j,t})^{2(1-\gamma)}\{1+2(1-\gamma)r^f \\
&\quad +(1-\gamma)\sigma^2(\alpha^*_{j,t})^2\}
\end{aligned}$$

We move onto the RHS of the Bellman equation:

$$RHS = \frac{c_{j,t}^{1-\gamma}}{1-\gamma}+\beta\phi_j[-\frac{1}{2}\theta A^2_{j+1}(w_{j,t}-c_{j,t})^{2-2\gamma}\sigma^2(\alpha^*_{j,t})^2+E_t[V^0_{j+1}(w^*_{j+1,t+1})]+\theta E_{j,t}[V^1_{j+1}(w^*_{j+1,t+1})]]$$

The FOC w.r.t $c_{j,t}$ gives:

$$\begin{aligned}
0 &= c_{j,t}^{-\gamma}-(w_{j,t}-c_{j,t})^{-\gamma}b_{j+1}^{-\gamma} \\
&\quad +\theta\beta\phi_j(w_{j,t}-c_{j,t})^{1-2\gamma}\{(A^2_{j+1}-B_{j+1})(1-\gamma)\sigma^2(\alpha^*_{j,t})^2-B_{j+1}[1+2(1-\gamma)r^f]\}
\end{aligned}$$

By plugging $c_{j,t}=c^*_{j,t}+\theta c'_{j,t}$, we get:

$$\gamma b_{j+1}^{-\gamma-1}c'_{j,t}+\gamma b_{j+1}^{-\gamma}c'_{j,t} = \beta\phi_j(w_{j,t}-c^*_{j,t})^{2-\gamma}[(A^2_{j+1}-B_{j+1})(1-\gamma)\sigma^2(\alpha^*_{j,t})^2-B_{j+1}[1+2(1-\gamma)r^f]]$$

Therefore:

$$\begin{aligned}
c'_{j,t} &= \frac{\beta\phi_j\{(A^2_{j+1}-B_{j+1})(1-\gamma)\sigma^2(\alpha^*_{j,t})^2-B_{j+1}[1+2(1-\gamma)r^f]\}}{\gamma b_{j+1}^{-\gamma}(1+b_{j+1}^{-1})}(w_{j,t}-c^*_{j,t})^{2-\gamma} \\
&= \frac{\beta\phi_j\{(A^2_{j+1}-B_{j+1})(1-\gamma)\sigma^2(\alpha^*_{j,t})^2-B_{j+1}[1+2(1-\gamma)r^f]\}}{\gamma b_{j+1}^{-\gamma}(1+b_{j+1}^{-1})}(\frac{w_{j,t}}{1+b_{j+1}})^{2-\gamma} \qquad (133)
\end{aligned}$$

where we use first-order Taylor approximation:

$$\begin{aligned}
c_{j,t}^{-\gamma} &\approx (c^*_{j,t})^{-\gamma}-\gamma(c^*_{j,t})^{-\gamma-1}\theta c'_{j,t} \\
(w_{j,t}-c_{j,t})^{-\gamma} &\approx (w_{j,t}-c^*_{j,t})^{-\gamma}+\gamma(w_{j,t}-c^*_{j,t})^{-\gamma-1}\theta c'_{j,t}
\end{aligned}$$

Plugging the above results into the Bellman equation gives:

$$\begin{aligned}
B_j\frac{w_{j,t}^{2(1-\gamma)}}{2(1-\gamma)} &= [\frac{1}{\gamma(1+b_{j+1}^{-1})}-\frac{1}{2(1-\gamma)}]\beta\phi_j\{(A^2_{j+1}-B_{j+1})(1-\gamma)\sigma^2(\alpha^*_{j,t})^2 \\
&\quad -B_{j+1}[1+2(1-\gamma)r^f]\}(w_{j,t}-c^*_{j,t})^{2-2\gamma}
\end{aligned}$$

where we use first-order Taylor approximation again:

$$c_{j,t}^{1-\gamma} \approx (c_{j,t}^*)^{1-\gamma} - (\gamma-1)(c_{j,t}^*)^{-\gamma}\theta c_{j,t}'$$

Matching coefficients w.r.t. $w_{j,t}$ gives:

$$B_j = \beta\phi_j[\frac{2(1-\gamma)}{\gamma(1+b_{j+1}^{-1})}-1]\{(A_{j+1}^2-B_{j+1})(1-\gamma)\sigma^2(\alpha_{j,t}^*)^2-B_{j+1}[1+2(1-\gamma)r^f]\}(\frac{1}{1+b_{j+1}})^{2-2\gamma} \tag{134}$$

with $B_J = 0$ because the agents in the last period do not invest and hence are not exposed to uncertainty. The system can be solved backward.

We can express $\alpha_{j,t}$ and $c_{j,t}$ in another way:

$$\alpha_{j,t} = \alpha^* + \theta\alpha^* w_{j,t}^{1-\gamma}\Omega_{\alpha j} \tag{135}$$

$$c_{j,t} = c_{j,t}^* + \theta w_{j,t}^{2-\gamma}\Omega_{cj} \tag{136}$$

where

$$\Omega_{\alpha j} = -\frac{A_{j+1}^2 + B_{j+1}(\gamma-1)}{\gamma A_{j+1}(1+b_{j+1})^{\gamma-1}} \tag{137}$$

$$\Omega_{cj} = \frac{\beta\phi_j\{(A_{j+1}^2 - B_{j+1})(1-\gamma)\sigma^2(\alpha^*)^2 - B_{j+1}[1+2(1-\gamma)r^f]\}}{\gamma b_{j+1}^{-\gamma-1}(1+b_{j+1})^{3-\gamma}} \tag{138}$$

For Section 3.3, we note that Ω_{cj} is potentially non-monotonic in ϕ_1, depending on parameters:

$$sgn(\frac{\partial\Omega_{c1}}{\partial\phi_1}) = sgn((1-\gamma)\frac{\partial\phi_1^{-\frac{1}{\gamma}}(1+\tilde{b}\phi_1^{-\frac{1}{\gamma}})^{\gamma-3}}{\partial\phi_1}) \tag{139}$$

$$= sgn((\gamma-1)(1+(\tilde{b}+\gamma-3)\phi_1^{-\frac{1}{\gamma}})) \tag{140}$$

Proposition 3: The result follows largely from applying Proposition 1.

Since $J = 2$, we have that $A_2 = 1$, and so:

$$b_2 = \left[\beta\phi_1[1+(1-\gamma)r^f + \frac{1}{2}(1-\gamma)\frac{(\mu-r^f)^2}{\gamma\sigma^2}]\right]^{-\frac{1}{\gamma}} \tag{141}$$

$$= \phi_1^{-\frac{1}{\gamma}}\tilde{b} \tag{142}$$

where $\tilde{b}$ is defined in (31).

Applying the definitions in Proposition 1, we further obtain:

$$A_1 = \left(\frac{\phi_1^{\frac{1}{\gamma}} + \tilde{b}}{\tilde{b}}\right)^{\gamma} \tag{143}$$

$$b_1 = \frac{\tilde{b}}{\phi_1^{\frac{1}{\gamma}} + \tilde{b}} \tag{144}$$

Substituting these into equation (18) gives equations (29) and (30). Finally, note using equations (10) and (9) that:

$$c_{2,t+2} = w_{2,t+2} = w_{0,t}R^p_{0,t+1}R^p_{1,t+2} - c_{0,t}R^p_{0,t+1}R^p_{1,t+2} - c_{1,t+1}R^p_{1,t+2} \tag{145}$$

Substituting in equations (29) and (30) implies equation (32).

Proposition 4: Since in a simple three-period model, the old do not invest in the stock market, we have $\nu_{2,t} = 0$. Substituting equations (27) and (28) into equation (24), and using the expressions for A_1, b_1 and b_2 gives us the optimal distortions for the young and the middle-aged.

Corollary 1: Using Proposition 4 we have that:

$$\nu_{1,t}(1) = -\frac{\theta\sigma\alpha^*}{(1+\tilde{b})^{1-\gamma}} w_{1,t}^{1-\gamma} \tag{146}$$

And:

$$\nu_{0,t}(0) = -\frac{\theta\sigma\alpha^*}{\tilde{b}^{\gamma-1}(\tilde{b}+\tilde{b}^2)^{1-\gamma}} w_{0,t}^{1-\gamma} = -\frac{\theta\sigma\alpha^*}{(1+\tilde{b})^{1-\gamma}} w_{0,t}^{1-\gamma} \tag{147}$$

Therefore $\nu_{0,t}(0) = \nu_{1,t}(1)$

Corollary 2: For the middle-aged:

$$\begin{aligned}
\frac{\partial \nu_{1,t}}{\partial \phi_1} &= -\theta\sigma\alpha^*_{1,t}(w_{1,t})^{1-\gamma}\frac{\partial(1+\phi_1^{-\frac{1}{\gamma}}\tilde{b})^{\gamma-1}}{\partial\phi} \\
&= \frac{\gamma-1}{\gamma}\theta\sigma\alpha^*_1(w_{1,t})^{1-\gamma}(1+\phi_1^{-\frac{1}{\gamma}}\tilde{b})^{\gamma-2}\phi_1^{-\frac{1}{\gamma}-1}\tilde{b}\begin{cases} <0 & \text{if } \gamma<1 \\ =0 & \text{if } \gamma=1 \\ >0 & \text{if } \gamma>1 \end{cases}
\end{aligned} \tag{148}$$

For the young:

$$\frac{\partial \nu_{0,t}}{\partial \phi_1} = -\theta\sigma\alpha^*_{0,t}(w_{0,t})^{1-\gamma}\tilde{b}^{-\gamma}(\phi_1^{\frac{1}{\gamma}} + \tilde{b} + \tilde{b}^2)^{-2+\gamma}\frac{1}{\gamma}\phi_1^{\frac{1}{\gamma}-1}(\tilde{b}^2 + \gamma\phi_1^{\frac{1}{\gamma}} + \tilde{b}\gamma) < 0 \tag{149}$$

Corollary 3: For the middle-aged:

$$\frac{\partial \nu_{1,t}}{\partial w_{1,t}} = -\frac{\theta\sigma\alpha^*\phi_1^{\frac{1-\gamma}{\gamma}}}{(\phi_1^{\frac{1}{\gamma}} + \tilde{b})^{1-\gamma}} w_{1,t}^{-\gamma}(1-\gamma) \begin{cases} < 0 & \text{if } \gamma < 1 \\ = 0 & \text{if } \gamma = 1 \\ > 0 & \text{if } \gamma > 1 \end{cases}$$

For the young:

$$\frac{\partial \nu_{0,t}}{\partial w_{0,t}} = -\frac{\theta\sigma\alpha^*(\phi_1^{\frac{1}{\gamma}} + \tilde{b})}{\tilde{b}^{\gamma}(\phi_1^{\frac{1}{\gamma}} + \tilde{b} + \tilde{b}^2)^{1-\gamma}} w_{0,t}^{-\gamma}(1-\gamma) \begin{cases} < 0 & \text{if } \gamma < 1 \\ = 0 & \text{if } \gamma = 1 \\ > 0 & \text{if } \gamma > 1 \end{cases}$$

Corollary 4: Differentiating equations (148) and (149) with respect to w_0, w_1 respectively:

$$\begin{aligned}
\frac{\partial}{\partial w_{0,t}}\left(\frac{\partial \nu_{0,t}}{\partial \phi_1}\right) &= -\frac{(1-\gamma)}{\gamma}\theta\sigma\alpha^*_{0,t}(w_{0,t})^{-\gamma}\tilde{b}^{-\gamma}(\phi_1^{\frac{1}{\gamma}} + \tilde{b} + \tilde{b}^2)^{-2+\gamma}\phi_1^{\frac{1}{\gamma}-1}(\tilde{b}^2 + \gamma\phi_1^{\frac{1}{\gamma}} + \tilde{b}\gamma) \\
\frac{\partial}{\partial w_{1,t}}\left(\frac{d\nu_{1,t}}{d\phi_1}\right) &= -\frac{(1-\gamma)^2}{\gamma}\theta\sigma\alpha^*_1(w_{1,t})^{-\gamma}(1 + \phi_1^{-\frac{1}{\gamma}}\tilde{b})^{\gamma-2}\phi_1^{-\frac{1}{\gamma}-1}\tilde{b}
\end{aligned}$$

If $\gamma = 1$, both differentials equal 0. If $\gamma \neq 0$, then:

$$\begin{aligned}
\frac{\partial}{\partial w_{0,t}}\left(\frac{\partial \nu_{0,t}}{\partial \phi_1}\right) &\propto -(1-\gamma) \\
\frac{\partial}{\partial w_{1,t}}\left(\frac{\partial \nu_{1,t}}{\partial \phi_1}\right) &\propto -(1-\gamma)^2
\end{aligned}$$

Combined with the signs derived in Corollary 2 this delivers Corollary 4.

Lemma 2: To reduce notation, in this proof we use $AD_{r,t}$ to denote the relative aggregate demand for risky assets:

$$AD_{r,t} \equiv \frac{AD_t(risky)}{AD_t(safe) + AD_t(risky)} \tag{150}$$

Using equation (17) to substitute out for α^*, equation (53) can be rewritten:

$$AD_{r,t} = \frac{\mu - r^f}{\gamma\sigma^2}\left(1 + \theta(\Gamma_{0,t} + \Gamma_{1,t})\right) \tag{151}$$

where:

$$\Gamma_{j,t} = \frac{\Omega_{\alpha,j} w_{j,t}^{1-\gamma}(w_{j,t} - c_{j,t})}{w_{0,t} - c_{0,t} + w_{1,t} - c_{1,t}} \tag{152}$$

for $j \in \{0, 1\}$.

From this we have:

$$\frac{\partial AD_{r,t}}{\partial \mu} = \frac{1}{\gamma\sigma^2}\left(1 + \theta(\Gamma_{0,t} + \Gamma_{1,t})\right) + \frac{\mu - r^f}{\gamma\sigma^2}\theta\left(\frac{\partial \Gamma_{0,t}}{\partial \mu} + \frac{\partial \Gamma_{1,t}}{\partial \mu}\right) \tag{153}$$

This derivative is strictly positive if:

$$-\theta\left(\frac{\partial \Gamma_{0,t}}{\partial \mu} + \frac{\partial \Gamma_{1,t}}{\partial \mu}\right) < AD_{r,t}\frac{\gamma\sigma^2}{(\mu - r^f)^2} \tag{154}$$

Similarly, differentiating equation (151) with respect to r^f yields:

$$\frac{\partial AD_{r,t}}{\partial r^f} = -\frac{1}{\gamma\sigma^2}\left(1 + \theta(\Gamma_{0,t} + \Gamma_{1,t})\right) + \frac{\mu - r^f}{\gamma\sigma^2}\theta\left(\frac{\partial \Gamma_{0,t}}{\partial r^f} + \frac{\partial \Gamma_{1,t}}{\partial r^f}\right) \tag{155}$$

This derivative is strictly negative if:

$$\theta\left(\frac{\partial \Gamma_{0,t}}{\partial r^f} + \frac{\partial \Gamma_{1,t}}{\partial r^f}\right) < AD_{r,t}\frac{\gamma\sigma^2}{(\mu - r^f)^2} \tag{156}$$

The right hand side is the same in conditions (154) and (156), and it is strictly positive. Let us start with condition (154). It is trivially satisfied if:

$$\left(\frac{\partial \Gamma_{0,t}}{\partial \mu} + \frac{\partial \Gamma_{1,t}}{\partial \mu}\right) \geq 0 \tag{157}$$

Outside of this case, condition (154) is satisfied if:

$$\theta < AD_{r,t} \frac{\gamma\sigma^2}{(\mu - r^f)^2} \left(-\frac{\partial \Gamma_{0,t}}{\partial \mu} - \frac{\partial \Gamma_{1,t}}{\partial \mu} \right)^{-1} \tag{158}$$

Now consider condition (156). Again, this is trivially satisfied if:

$$\left(\frac{\partial \Gamma_{0,t}}{\partial r^f} + \frac{\partial \Gamma_{1,t}}{\partial r^f} \right) \leq 0 \tag{159}$$

Outside of this case, condition (156) is satisfied if:

$$\theta < AD_{r,t} \frac{\gamma\sigma^2}{(\mu - r^f)^2} \left(\frac{\partial \Gamma_{0,t}}{\partial r^f} + \frac{\partial \Gamma_{1,t}}{\partial r^f} \right)^{-1} \tag{160}$$

A sufficient condition for both (154) and (156) to be satisfied is therefore:

$$\theta < \theta^* = AD_{r,t} \frac{\gamma\sigma^2}{(\mu - r^f)^2} \cdot \min \left(\left(-\frac{\partial \Gamma_{0,t}}{\partial \mu} - \frac{\partial \Gamma_{1,t}}{\partial \mu} \right)^{-1}, \left(\frac{\partial \Gamma_{0,t}}{\partial r^f} + \frac{\partial \Gamma_{1,t}}{\partial r^f} \right)^{-1} \right) \tag{161}$$

B Consumption and Saving with Ambiguity Aversion

In the benchmark without ambiguity, Proposition 3 implies that young and middle-aged agents consume less, and save more, when the probability of surviving to old age increases:

$$\frac{dc^*_{j,t}}{d\phi_1} < 0 \tag{162}$$

With ambiguity, we start with equation (28) for $j = \{0, 1\}$, and substitute out for $c^*_{j,t}$ using Proposition 3:

$$c_{0,t} = \frac{\tilde{b}^2}{\phi_1^{\frac{1}{\gamma}} + \tilde{b} + \tilde{b}^2} w_{0,t} + \theta w_{0,t}^{2-\gamma} \Omega_{c,0} \tag{163}$$

$$c_{1,t+1} = \frac{\tilde{b}}{\phi_1^{\frac{1}{\gamma}} + \tilde{b} + \tilde{b}^2} R_0^p w_{0,t} + \theta w_{1,t+1}^{2-\gamma} \Omega_{c,1} \tag{164}$$

Differentiating with respect to ϕ_1, we obtain:

$$\frac{dc_{0,t}}{d\phi_1} = -\frac{\tilde{b}^2\phi_1^{\frac{1-\gamma}{\gamma}}}{\gamma(\phi_1^{\frac{1}{\gamma}} + \tilde{b} + \tilde{b}^2)} w_{0,t} + \theta w_{0,t}^{2-\gamma}\frac{d\Omega_{c,0}}{d\phi_1} \tag{165}$$

$$\frac{dc_{1,t+1}}{d\phi_1} = -\frac{\tilde{b}\phi_1^{\frac{1-\gamma}{\gamma}}}{\gamma(\phi_1^{\frac{1}{\gamma}} + \tilde{b} + \tilde{b}^2)} R_0^p w_{0,t} + \theta\left((2-\gamma)w_{1,t+1}^{1-\gamma}\Omega_{c,1}\frac{dw_{1,t+1}}{d\phi_1} + w_{1,t+1}^{2-\gamma}\frac{d\Omega_{c,1}}{d\phi_1}\right) \tag{166}$$

These derivatives are negative whenever:

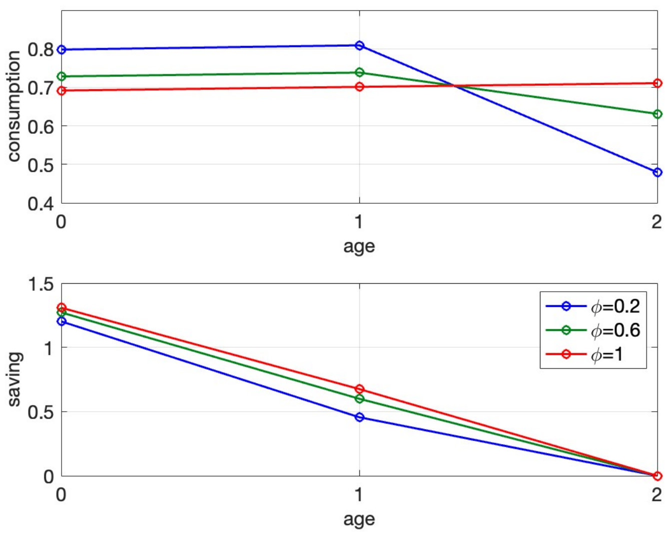
$$\theta\frac{d\Omega_{c,0}}{d\phi_1} < \frac{\tilde{b}^2\phi_1^{\frac{1-\gamma}{\gamma}}}{\gamma(\phi_1^{\frac{1}{\gamma}} + \tilde{b} + \tilde{b}^2)} w_{0,t}^{\gamma-1} \tag{167}$$

$$\theta\left((2-\gamma)\Omega_{c,1}\frac{dw_{1,t+1}}{d\phi_1} + w_{1,t+1}\frac{d\Omega_{c,1}}{d\phi_1}\right) < \frac{\tilde{b}\phi_1^{\frac{1-\gamma}{\gamma}}}{\gamma(\phi_1^{\frac{1}{\gamma}} + \tilde{b} + \tilde{b}^2)} R_0^p w_{0,t} w_{1,t+1}^{\gamma-1} \tag{168}$$

In both of these inequalities, the right hand side is strictly positive. Since $\Omega_{c,j}$ is independent of θ, there is therefore a θ^+ such that $\theta < \theta^+$ is sufficient to ensure both inequalities hold, and consumption of young and middle-aged agents falls as ϕ_1 rises.

Figure 23 shows a numerical example of this. It plots the consumption and saving paths for the ambiguity-averse agents with $J = 2$ studied in Section 3.2, with the same parameters as those in Figures 2 and 3.

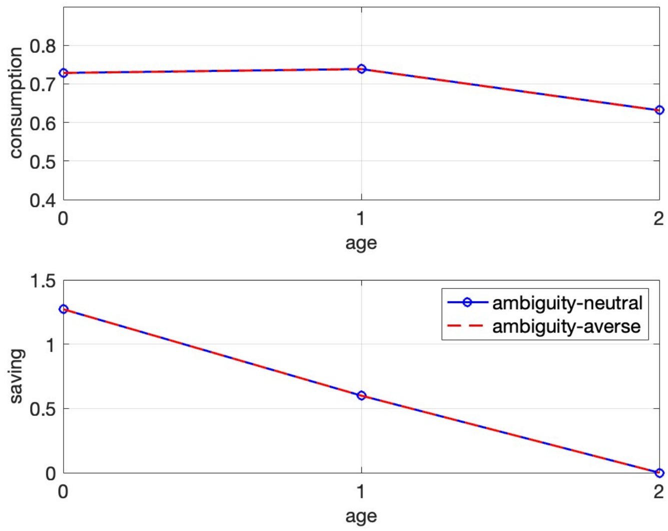
Figure 23: Consumption and saving paths with ambiguity.



Note: Plots constructed using $J = 2$, $\mu = 0.06$, $r^f = 0.045$, $\sigma = 0.1$, $\theta = 0.045$, $\gamma = 3$, $\beta = 0.99$, $\phi_1 \in (0, 1]$, $w_{0,t} = 2$, and risky asset returns set to their expected level every period. This therefore abstracts from the effect of return shocks.

These paths are very similar to those without ambiguity aversion in Figure 1. Figure 24 plots the paths of consumption for $\phi_1 = 0.6$, with and without ambiguity.

Figure 24: Consumption and saving paths with and without ambiguity, for $\phi_1 = 0.6$.



Note: Plots constructed using $J = 2$, $\mu = 0.06$, $r^f = 0.045$, $\sigma = 0.1$, $\phi_1 = 0.6$, $\gamma = 3$, $\beta = 0.99$, $\phi_1 \in (0, 1]$, $w_{0,t} = 2$, and risky asset returns set to their expected level every period. This therefore abstracts from the effect of return shocks.

C Section 4 Robustness Tests

Table 19 repeats the analysis of Table 2, this time replacing the basic (unconditional) equity share measure with the transformed measure $\tilde{\alpha}^*_{it}$, defined as:

$$\tilde{\alpha}^*_{it} = \frac{\tilde{\alpha}_{it} \cdot \text{financial wealth}}{\text{total net wealth exc. retirement accounts}} \tag{169}$$

Since the original survey measure $\tilde{\alpha}_{it}$ gives the share of (non-retirement fund) financial wealth invested in stocks, the numerator of the definition of $\tilde{\alpha}^*_{it}$ is the total amount invested in stocks. Dividing that by net wealth gives a measure of the share of overall assets invested in stocks. We exclude retirement accounts (IRAs and 401ks) from this calculation as these may be partly invested in stocks, but we are unable to observe the proportions. The quantitative and qualitative results are as in Table 2.

Tables 20 and 21 then repeat the same analysis using the subjective probability of surviving to age 65 and 75 respectively. The results are qualitatively similar to Table 2, but less precisely estimated because of the smaller variation in these subjective survival probabilities.

Table 19: Regressions on share of total assets invested in equity.

	(1) Equity share	(2) Equity share	(3) Equity share
Age	-0.0922	-0.1168	-0.1164
	(0.0815)	(0.0862)	(0.0919)
Subjective survival prob.	-0.1159*	-0.1221**	-0.1018
	(0.0597)	(0.0612)	(0.0639)
Age × Subjective survival prob.	0.0030**	0.0032**	0.0030*
	(0.0015)	(0.0016)	(0.0016)
Net wealth (\$1000s)		0.0002	0.0000
		(0.0001)	(0.0001)
Income (\$1000s)		0.0068	0.0064
		(0.0056)	(0.0056)
Period FE	Yes	Yes	Yes
Demographic controls	No	No	Yes
Observations	3440	3422	3414
R^2	0.0021	0.0084	0.0285

Note: Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *For other regression details see the note to Table 2.*

Table 22 repeats columns 2 and 3 of Table 2 with income and net wealth included in logs, rather than in levels. This entails removing the minority of observations with zero or negative values for these variables. The results are qualitatively and quantitatively robust to this change.

Finally, Table 23 repeats Table 2, but with the estimation conducted after winsorizing the subjective survival probability measure. Specifically, to ensure extreme survival probabilities are not driving our results, we drop all households with a subjective survival probability $\tilde{\phi}_{it} \leq 5\%$ (5% of the sample) and with $\tilde{\phi}_{it} \geq 95\%$ (7% of the sample). If anything, our results are stronger after removing these extreme values.

Table 20: Regressions on share of financial assets invested in equity, using subjective survival probability to age 65.

	(1) Equity share	(2) Equity share	(3) Equity share
Age	-0.1015	-0.1723	-0.1489
	(0.2126)	(0.2078)	(0.2046)
Subjective survival prob.	-0.0675	-0.0858	-0.1369
	(0.1104)	(0.1093)	(0.1053)
Age $\times$ Subjective survival prob.	0.0026	0.0031	0.0031
	(0.0025)	(0.0024)	(0.0024)
Net wealth ($1000s)		0.0007*	0.0004
		(0.0004)	(0.0003)
Income ($1000s)		0.0024	0.0018
		(0.0017)	(0.0013)
Period FE	Yes	Yes	Yes
Demographic controls	No	No	Yes
Observations	3493	3473	3465
R^2	0.0049	0.0110	0.0733

Note: Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *For other regression details see the note to Table 2.*

Table 21: Regressions on share of financial assets invested in equity, using subjective survival probability to age 75.

	(1) Equity share	(2) Equity share	(3) Equity share
Age	-0.0043	-0.0409	-0.0286
	(0.1406)	(0.1396)	(0.1349)
Subjective survival prob.	-0.0394	-0.0420	-0.0762
	(0.0847)	(0.0848)	(0.0796)
Age × Subjective survival prob.	0.0019	0.0019	0.0020
	(0.0019)	(0.0019)	(0.0018)
Net wealth ($1000s)		0.0007*	0.0004
		(0.0004)	(0.0003)
Income ($1000s)		0.0023	0.0018
		(0.0017)	(0.0013)
Period FE	Yes	Yes	Yes
Demographic controls	No	No	Yes
Observations	3487	3467	3459
R^2	0.0053	0.0111	0.0729

Note: Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *For other regression details see the note to Table 2.*

Table 22: Regressions on share of financial assets invested in equity, with income and net wealth in logs.

	(1) Equity share	(2) Equity share
Age	-0.2982***	-0.2082*
	(0.1122)	(0.1148)
Subjective survival prob.	-0.1942**	-0.1849**
	(0.0850)	(0.0843)
Age $\times$ Subjective survival prob.	0.0044**	0.0042**
	(0.0018)	(0.0018)
Log net wealth	3.6744***	3.2447***
	(0.4317)	(0.4789)
Log income	1.8458***	1.1855**
	(0.5838)	(0.5902)
Period FE	Yes	Yes
Demographic controls	No	Yes
Observations	2930	2924
R^2	0.0627	0.1001

Note: Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *For other regression details see the note to Table 2.*

Table 23: Regressions on share of financial assets invested in equity, after winsorizing survival probability.

	(1) Equity share	(2) Equity share	(3) Equity share
Age	-0.1956	-0.2008	-0.1870
	(0.1243)	(0.1235)	(0.1249)
Subjective survival prob.	-0.2490***	-0.2294**	-0.2288**
	(0.0965)	(0.0964)	(0.0957)
Age × Subjective survival prob.	0.0060***	0.0055**	0.0054**
	(0.0022)	(0.0022)	(0.0021)
Net wealth ($1000s)		0.0007*	0.0004
		(0.0004)	(0.0003)
Income ($1000s)		0.0064*	0.0044*
		(0.0034)	(0.0025)
Period FE	Yes	Yes	Yes
Demographic controls	No	No	Yes
Observations	3052	3034	3028
R^2	0.0064	0.0154	0.0791

Note: Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *For other regression details see the note to Table 2.*

D Quantitative Model Details

D.1 Model Description

Demographics, Assets, Preferences: Demographics are as in Section 2.1. We also maintain the assumption that there are two assets available for agents to invest in, one safe and one risky. The returns on these assets are as in Section 2.1.

An agent of age j chooses their consumption and portfolio allocation to maximize expected discounted lifetime utility:[33]

$$V_j^\theta(w_{j,t}, P_{j,t}) = \max_{c_{j,t},\tilde{\alpha}_{j,t}} \min_{\nu_{j,t}} \left\{ \frac{c_{j,t}^{1-\gamma}}{1-\gamma} - \beta \left(E_{j,t}\left[-\phi_j \left(\frac{1}{2\theta_{j,t}} \nu_{j,t}^2 + V_{j+1}^\theta(w_{j+1,t+1}, P_{j+1,t+1}) \right) \right]^{1-\chi} \right)^{\frac{1}{1-\chi}} \right\} \tag{170}$$

This is as in equation (21), except that there is an extra state variable (permanent income $P_{j,t}$, defined below) and agents have Epstein-Zin preferences whenever the value of χ differs from 0. This formulation of Epstein-Zin preferences is as in Rudebusch and Swanson (2012), and is equivalent to the form in e.g. Gomes and Michaelides (2005). Here the EIS is equal to γ^{-1}, and the coefficient of relative risk aversion is $1-(1-\chi)(1-\gamma)$. We also allow the ambiguity aversion parameter $\theta_{j,t}$ to vary over time in a specific way to permit a common normalization used in the model solution, described below.

For robustness exercises below, we also consider a version of this objective with a bequest motive, where the form of the bequest motive follows Gomes and Michaelides (2005):[34]

$$V_j^\theta(w_{j,t}, P_{j,t}) = \max_{c_{j,t},\tilde{\alpha}_{j,t}} \min_{\nu_{j,t}} \left\{ \frac{c_{j,t}^{1-\gamma}}{1-\gamma} - \beta \left(E_{j,t}\left[- \left(\frac{1}{2\theta_{j,t}} \nu_{j,t}^2 + \phi_j V_{j+1}^\theta(w_{j+1,t+1}, P_{j+1,t+1}) + (1-\phi_j)\frac{b^\gamma}{1-\gamma} w_{j+1,t+1} \right) \right]^{1-\chi} \right)^{\frac{1}{1-\chi}} \right\} \tag{171}$$

[33]As in Section 4, $\tilde{\alpha}_{j,t}$ refers to the fraction of an agent's financial assets invested in the risky asset.

[34]Note that there is a small change in the specification of ambiguity aversion in this case, as the penalty to belief distortions is no longer multiplied by ϕ_j (i.e. the penalty is still paid if the agent dies). If we did not make this change, then distortions in the terminal period would be cost-free, but there would still be a benefit to distorting beliefs through the impact on the expected utility from bequests. The optimal belief distortion would therefore be infinite. To avoid this pathological solution, we therefore change the ambiguity specification accordingly for this robustness check.

Income and Wealth: Agents are born at age $j = 20$ with no assets. Until the age of 66, they are employed, earning risky labor income $y_{j,t}$ with the law of motion:

$$y_{j,t} = \underbrace{\exp[f_j + q_{j,t}]}_{\equiv P_{j,t}} \cdot \exp[u_{j,t}] \tag{172}$$

$$q_{j,t} = q_{j-1,t-1} + z_{j,t} \tag{173}$$

where f_j is a deterministic function of age, which we set to a cubic with parameters as in Cocco et al. (2005); $q_{j,t}$ is a persistent shock, $u_{j,t}$ and $z_{j,t}$ are transitory shocks with i.i.d. mean-zero Gaussian distributions, and $P_{j,t}$ is permanent income. A more detailed description of the income process is found in Gomes (2020) - equation 172 here reproduces his equation 18.

At age 67, agents retire. From then until their death they earn a risk-free retirement income, given by:

$$y_{j,t} = \exp[\lambda(f_K + q_{K,t_r})] \tag{174}$$

where retirement occured in period t_r at age K, and λ is a constant determining the generosity of retirement income. Log retirement income is therefore a constant fraction of the log of the persistent components of income in the final period of the agent's working life.

Wealth then accumulates according to:

$$w_{j+1,t+1} = \begin{cases} (R^f + (R_{t+1} - R^f + \sigma_1 \nu_{j,t})\tilde{\alpha}_{j,t}) \cdot (w_{j,t} - c_{j,t}) + y_{j+1,t+1} \cdot (1 - F) & \text{if } \tilde{\alpha}_{j,t} > 0 \\ R^f \cdot (w_{j,t} - c_{j,t}) + y_{j+1,t+1} & \text{if } \tilde{\alpha}_{j,t} = 0 \end{cases} \tag{175}$$

where F is a participation cost, set proportional to income to reflect that much of the cost is in terms of the agent's time. These processes for income and wealth follow the standard form in the portfolio choice literature (see e.g. the discussion in Gomes, 2020). Indeed, the only differences to Gomes and Michaelides (2005) in these processes are that there is no housing expenditure, and that equity market participation costs are paid per-period (as in e.g. Catherine, 2022) rather than once in an agent's lifetime. The only difference to Cocco et al. (2005) is that they do not include a participation cost.

Normalization: The model is solved with standard numerical techniques. As in Cocco et al. (2005), we take advantage of the fact that the utility function and constraints can be scaled by permanent income $P_{j,t}$. This removes one state variable from the model and so makes the solution substantially faster.

Specifically, define consumption, income, and wealth relative to permanent income as $\hat{c}_{j,t} = \frac{c_{j,t}}{P_{j,t}}$, $\hat{y}_{j,t} = \frac{y_{j,t}}{P_{j,t}}$, and $\hat{w}_{j,t} = \frac{w_{j,t}}{P_{j,t}}$. To enable this normalization, we also assume that $\theta_{j,t} = \frac{\hat{\theta}}{P_{j,t}^{1-\gamma}}$ for some constant $\hat{\theta}$.[35] With these definitions, the objective function and constraints can be written as:

$$\hat{V}_j^\theta(\hat{w}_{j,t}) = \max_{\hat{c}_{j,t},\tilde{\alpha}_{j,t}} \min_{\nu_{j,t}} \left\{ \frac{\hat{c}_{j,t}^{1-\gamma}}{1-\gamma} - \beta \left(E_{j,t}\left[-\phi_j \left(\frac{1}{2\hat{\theta}}\nu_{j,t}^2 + g_{j+1,t+1}^{1-\gamma}\hat{V}_{j+1}^\theta(\hat{w}_{j+1,t+1}) \right) \right]^{1-\chi} \right)^{\frac{1}{1-\chi}} \right\} \tag{176}$$

$$\hat{w}_{j+1,t+1} = \begin{cases} (R^f + (R_{t+1} - R^f + \sigma_1\nu_{j,t})\tilde{\alpha}_{j,t}) \cdot \frac{(\hat{w}_{j,t}-\hat{c}_{j,t})}{g_{j+1,t+1}} + \hat{y}_{j+1,t+1} \cdot (1-F) & \text{if } \tilde{\alpha}_{j,t} > 0 \\ R^f \cdot \frac{(\hat{w}_{j,t}-\hat{c}_{j,t})}{g_{j+1,t+1}} + \hat{y}_{j+1,t+1} & \text{if } \tilde{\alpha}_{j,t} = 0 \end{cases} \tag{177}$$

where $\hat{V}_j^\theta(\hat{w}_{j,t}) \equiv \frac{V_j^\theta(w_{j,t},P_{j,t})^{\frac{1}{1-\gamma}}}{P_{j,t}}$, and $g_{j+1,t+1} \equiv \frac{P_{j+1,t+1}}{P_{j,t}}$.

In the model with bequests, the constraint remains as above, but the objective function becomes:

$$\begin{aligned} \hat{V}_j^\theta(\hat{w}_{j,t}) = \max_{\hat{c}_{j,t},\tilde{\alpha}_{j,t}} \min_{\nu_{j,t}} \Bigg\{ \frac{\hat{c}_{j,t}^{1-\gamma}}{1-\gamma} - \beta \Bigg(E_{j,t}\Bigg[& \frac{1}{2\hat{\theta}}\nu_{j,t}^2 - \phi_j g_{j+1,t+1}^{1-\gamma}\hat{V}_{j+1}^\theta(\hat{w}_{j+1,t+1}) \\ & - (1-\phi_j)\frac{b^\gamma}{1-\gamma} g_{j+1,t+1}^{1-\gamma}\hat{w}_{j+1,t+1}^{1-\gamma} \Bigg]^{1-\chi} \Bigg)^{\frac{1}{1-\chi}} \Bigg\} \end{aligned} \tag{178}$$

Inner Minimization: To solve for agent decision rules, it is helpful to derive the first order condition for the optimal belief distortion $\nu_{j,t}$. This also provides useful intuition for the model results. The resulting first order condition is:

$$\nu_{j,t} = -\hat{\theta}\sigma_1\tilde{\alpha}_{j,t}(\hat{w}_{j,t} - \hat{c}_{j,t}) \cdot \frac{1}{E_{j,t}Z_{j,t}^{-\chi}} \cdot E_{j,t}Z_{j,t}^{-\chi} g_{j+1,t+1}^{-\gamma} \frac{\partial \hat{V}_{j+1}^\theta(\hat{w}_{j+1,t+1})}{\partial \hat{w}_{j+1,t+1}} \tag{179}$$

[35]This implies that the degree of ambiguity aversion is increasing in permanent income. The main consequences of ambiguity aversion identified in the simple model without this assumption continue to hold here.

where:

$$Z_{j,t} = -\phi_j \left(\frac{\nu_{j,t}^2}{2\hat{\theta}} + g_{j+1,t+1}^{1-\gamma} \hat{V}_{j+1}^{\theta}(\hat{w}_{j+1,t+1}) \right) \tag{180}$$

Notice that setting $\chi = 0$ (CRRA utility) and $g_{j+1,t+1} = 1$ (removing risky income growth) returns us to the form of the inner minimization FOC in the simple model (equation (41)). The same forces driving the reaction of belief distortions to changes in survival probabilities therefore operate in this richer quantitative model as in the simple model of Section 3.

In the case with a bequest motive we have:

$$\nu_{j,t} = -\hat{\theta}\sigma_1 \tilde{\alpha}_{j,t}(\hat{w}_{j,t} - \hat{c}_{j,t}) \cdot \frac{1}{E_{j,t}\tilde{Z}_{j,t}^{-\chi}} \cdot E_{j,t}\tilde{Z}_{j,t}^{-\chi} g_{j+1,t+1}^{-\gamma} \left[\phi_j \frac{\partial \hat{V}_{j+1}^{\theta}(\hat{w}_{j+1,t+1})}{\partial \hat{w}_{j+1,t+1}} + (1-\phi_j) b^{\gamma} \hat{w}_{j+1,t+1}^{-\gamma} \right] \tag{181}$$

where:

$$\tilde{Z}_{j,t} = -\left(\frac{\nu_{j,t}^2}{2\hat{\theta}} + \phi_j g_{j+1,t+1}^{1-\gamma} \hat{V}_{j+1}^{\theta}(\hat{w}_{j+1,t+1}) + (1-\phi_j) \frac{b^{\gamma}}{1-\gamma} g_{j+1,t+1}^{1-\gamma} \hat{w}_{j+1,t+1}^{1-\gamma} \right) \tag{182}$$

D.2 Data Construction and Calibration

Standard parameters: The majority of model parameters are taken from Gomes (2020), which is itself based on Cocco et al. (2005) and Gomes and Michaelides (2005). As in those papers, the age-dependent component of income is assumed to be given by:

$$f_j = aa + b1 \times j + b2 \times j^2 + b3 \times j^3 \tag{183}$$

These calibrated parameters are listed in Table 24. The only ones different to Gomes (2020) are γ and χ: where Gomes (2020) assumes CRRA utility for his baseline analysis (and so $\chi = 0$), we separate risk aversion and the elasticity of intertemporal substitution. We select γ and χ so that the EIS is 0.35 and the coefficient of relative risk aversion is 5. These are common values in this literature (see e.g. Foltyn, 2025).

Mortality: For the main calibration, we take mortality rates by age group from the Office of the Chief Actuary at the Social Security Administration, and compute ϕ_j as $1 - \text{mortality rate}_j$. For projections to 2100 in Section 5.3, we use projected mortality rates from the same data source. They project mortality rates for males and females separately, at every year of age up to 119. To compute aggregate survival rates from these series, we first take data on sex ratios in the US in 2021 from the UN World Population Prospects 2022. This data is reported at selected ages,[36] so for the ages missing from the sex ratio data we assume the sex ratio is equal to that of the nearest cohort for which there is data. We then generate projected sex ratios at each age for all years up to 2100, by combining the mortality rates from the Office of the Chief Actuary for each sex, age, and year, with the assumption that the sex rate at birth will remain the same as it was in 2021. These projections are important, as in 2019 the female-male ratio rose sharply at older age groups due to women having longer life expectancy in the US. However, this life expectancy gap is projected to narrow in coming decades, which will reduce the female-male ratio in older age groups. Finally, we use the projected sex ratios to combine female and male mortality rates to give an aggregate mortality rate. 1- this mortality rate for each age group then gives projected ϕ_j in 2100.

Targets: We choose the risky asset participation cost to match the average participation rate in the 2019 wave of the Survey of Consumer Finances, which was 0.53 (Bhutta et al., 2020). We choose θ to match β_2 and β_3 in column 3 of Table 3, as described in Section 5.1. This step involves choosing a value ζ that regulates the heterogeneity in subjective life expectancy in the simulated data used for the regressions. This is chosen to match the standard deviation of subjective probabilities of surviving to age 85 among respondents in the SCE who answered the question on their equity share (28%). These targets imply parameter values of $F = 0.036$, $\zeta = 0.88$, and $\theta = 0.19$.[37]

Table 24: Calibrated parameters.

Parameter	Value	Parameter	Value
aa	0.530339	γ	$\frac{20}{7}$
b1	0.16818	λ	0.68212
b2	-0.00323371	μ	1.055
b3	0.000019704	σ_1	0.2
R^f	1.015	σ_u^2	0.1
χ	$-\frac{15}{13}$	σ_z^2	0.1
δ	0.97		

Note: Calibrated parameters used in the baseline model, taken from Gomes (2020), with the exception of γ and χ, which are taken to generate an EIS of 0.35 and a coefficient of relative risk aversion of 5, as in Foltyn (2025).

[36] These ages are 15, 20, 30, 40, 50, 60, 70, 80, 90, 100.

[37] When evaluating the value of ζ, recall that annual mortality rates are typically small numbers. The mortality rate of 60-year-olds in 2019 was 0.90%. Our ζ calibration implies that the perceived mortality rate at 60 varies between 0.11% and 1.71%. This is why a large ζ value is required to generate the substantial heterogeneity in survival probabilities reported in the SCE data.

D.3 Detection Error Probabilities

As is common in the literature on ambiguity aversion, we use model detection error probabilities (DEP) to infer whether agents are hedging against the models that are empirically plausible to generate the data we observe. Intuitively, we treat agents as statisticians using likelihood ratio test to discriminate among models. DEP measures how far the alternative models can deviate from the approximating one without being discarded. Low values of DEP means that agents are unwilling to discard very different alternative models, which could be easily discriminated given observed data.

DEP assigns equal initial priors to the approximating and distorted models, hence it is the average of the probabilities of Type I and Type II errors. A Type I error occurs when the likelihood ratio test chooses the distorting model when the approximating model is the true data generating process. A Type II error is the reverse. Formally, DEP is defined as:

$$DEP = \frac{1}{2}Prob(\ln(\frac{L_A}{L_B}) < 0|A) + \frac{1}{2}Prob(\ln(\frac{L_B}{L_A}) > 0|B) \tag{184}$$

where A denotes the approximating model and B is the distorting model.

As discussed by Anderson, Hansen and Sargent (2003), the following bound on the average error in using a likelihood ratio test to discriminate between the approximating and distorted models is useful when the data is of a continuous record with length T. The DEP bound in this discrete-time model can be approximated in the following way:

$$avg\ DEP \leq \frac{1}{2}E\ exp\{-\frac{1}{8}\int_{t=0}^{T}\nu^2(w_t)\,dt\} \tag{185}$$

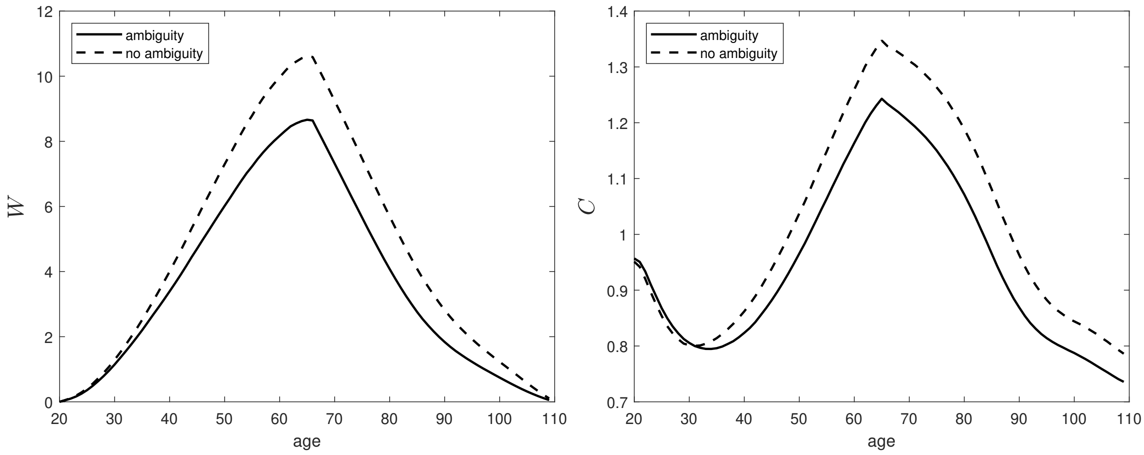
$$\approx \frac{1}{2}E\ exp\{-\frac{1}{8}\sum_{j=20}^{J}\nu^2(w_j)\} \tag{186}$$

In the calibration, we take $J = 109$, which is consistent with the maximum lifespan in our calibration, and set the initial age to $j = 20$. With the calibrated value of $\theta = 0.19$, we obtain a large average DEP value of 0.487. Besides, if instead we use the distortion for the poorest individual in each age group, we obtain a DEP value of 0.462, which acts as a lower bound for all agents. This implies that based on the observed data, it is not easy to distinguish the distorted model from the approximating model.

D.4 Other life cycle profiles

Figure 25 shows the average wealth and consumption over agent life-cycles in the model, with and without ambiguity. Both variables are expressed relative to permanent income, as in Gomes (2020) Figure 3a.[38] As in the simple model, ambiguity does not affect the qualitative shape of these profiles. The main effect of ambiguity is to reduce risky asset shares, particularly for younger agents (Figure 8), reducing average portfolio returns, and thus limiting wealth accumulation somewhat. To compensate for this lower wealth accumulation, agents consume less than in the no-ambiguity case.

Figure 25: Further life-cycle profiles I.



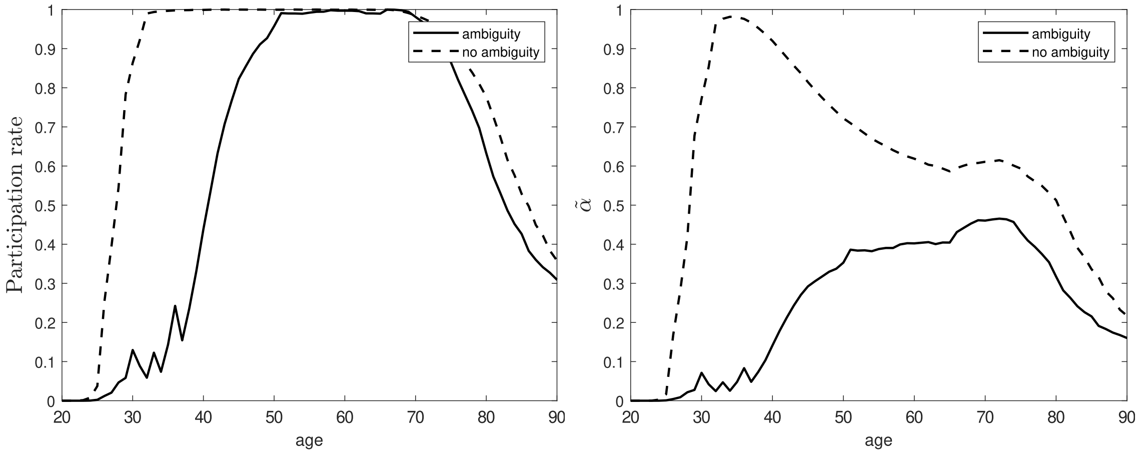
Note: Plots constructed using the calibration and data described in Section 5.1 and Appendix D.2. All variables are expressed after normalizing by permanent income $P_{j,t}$.

Figure 26 shows the risky asset participation rate, and the average unconditional risky asset share, over the life-cycle. The participation rate inherits the hump-shaped profile from the no-ambiguity model. Ambiguity does however substantially delay the rise in participation at young ages. Empirically, participation rates in the 2019 SCF display a similar hump-shaped profile, with a peak at age 45-54. Studying the SCF from 1998-2007, Chang et al. (2018) find a hump-shaped profile with a peak around age 60. While we therefore match the approximate location of peak participation (unlike the model without ambiguity), the hump-shape in our model is too pronounced relative to the data, where the maximum participation rate is typically around 60%.

As with the conditional risky asset share in Figure 8, ambiguity lowers the unconditional risky asset share, especially for young agents. It also makes the profile upward-sloping for most of the life-cycle. This upward slope is consistent with the SCE data (see Figure 6), and the overall profile is qualitatively consistent with estimates from the SCF – which similarly finds a gradual upward slope, with a decline in later years (see Chang et al., 2018, Figure 1B).

[38]The initial decline in consumption therefore occurs because permanent income grows faster than consumption at early ages, not because the level of consumption declines. The same path is present in Gomes (2020) Figure 3a.

Figure 26: Further life-cycle profiles II.



Note: Plots constructed using the calibration and data described in Section 5.1 and Appendix D.2.

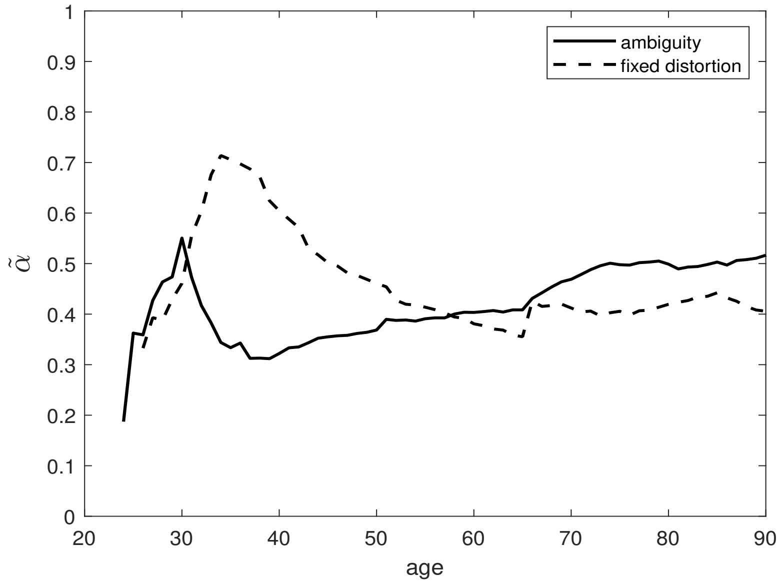
Figure 27 repeats Figure 8, but replaces the 'no-ambiguity' line with another alternative that highlights more explicitly how changing $\nu_{j,t}$ over the life-cycle affects the results. Specifically, the comparison is to a model in which $\nu_{j,t}$ is fixed to be -0.0663 for all agents in all periods. This ensures that the average conditional risky asset share is the same in both cases plotted.[39] This 'constant-distortion' case has the same qualitative shape as the no-ambiguity case in Figure 8, highlighting that the key results discussed in Section 5.2 are driven by the changes in $\nu_{j,t}$ over the life-cycle, and not by the mere presence of belief distortions.

D.5 Results with a bequest motive

Figure 28 reproduces Figure 10 with a bequest motive in utility, as in equation (171). The bequest parameter b is set to 2.5, as in Gomes and Michaelides (2005). Although the overall life-cycle profile is less strongly upward-sloping that in Figure 10, especially at older ages, the key qualitative results remain the same. The life-cycle profile still slopes upwards from early middle age onward, and increased longevity makes the gradient steeper.

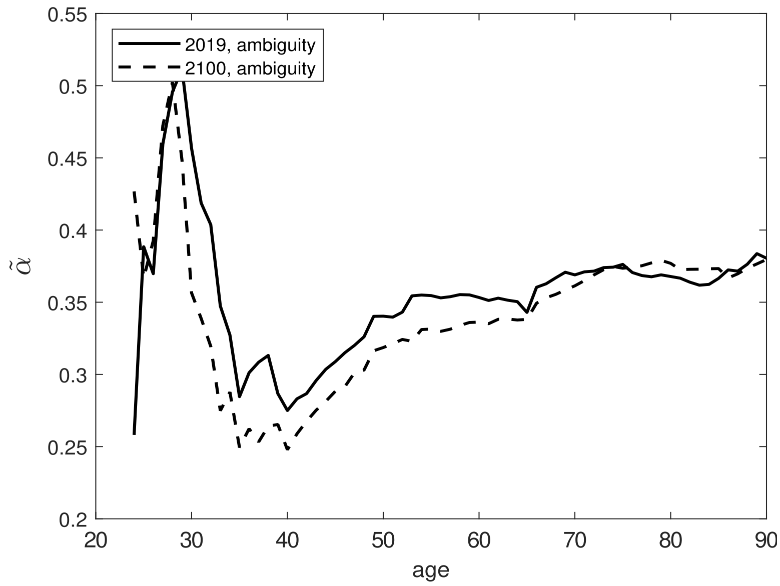
[39]Specifically, the constant level of $\nu_{j,t}$ is chosen to match the unweighted mean of the points in the age-profile plotted for the baseline case. That is, we do not account for the fact that there are fewer agents present in older age groups as some die before reaching them. This ensures that the average levels of the two lines in the graph are the same, giving the clearest comparison.

Figure 27: Model-implied conditional risky asset shares, 2019, vs the fixed-ν alternative.



Note: Plots constructed using the calibration and data described in Section 5.1 and Appendix D.2, with the 'fixed distortion' line constructed under the assumption that $\nu_{j,t} = -0.0663$ for all agents in all periods.

Figure 28: Model-implied conditional risky asset shares, 2019 and 2100, with a bequest motive.

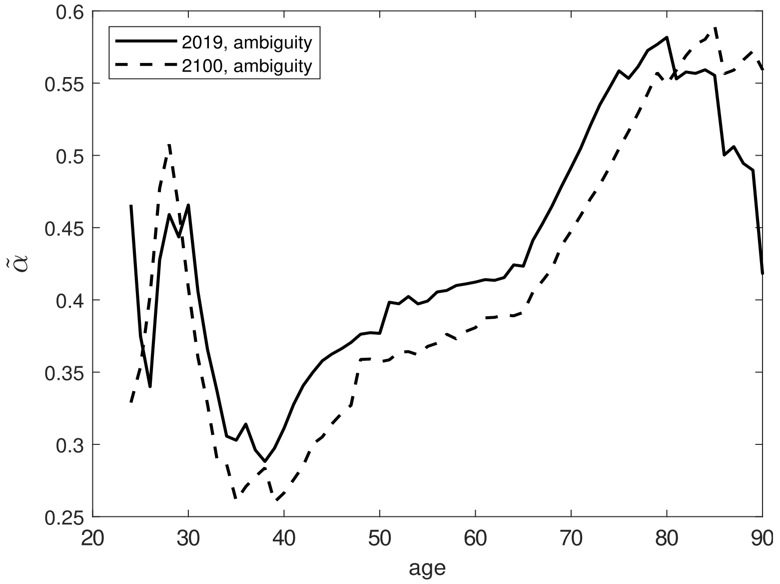


Note: Plots constructed using the calibration and data described in Section 5.1 and Appendix D.2, with a bequest motive in utility as described in Appendix D.1.

D.6 Results with homogeneous mortality rates

Figure 29 reproduces Figure 10 with homogeneous mortality rates. All agents have accurate beliefs about their mortality rates, set to the US data described above. The key qualitative results remain the same. The life-cycle profile is upward sloping from early middle age onward, and increased longevity makes the gradient steeper.

Figure 29: Model-implied conditional risky asset shares, 2019 and 2100, with homogeneous perceived mortality.



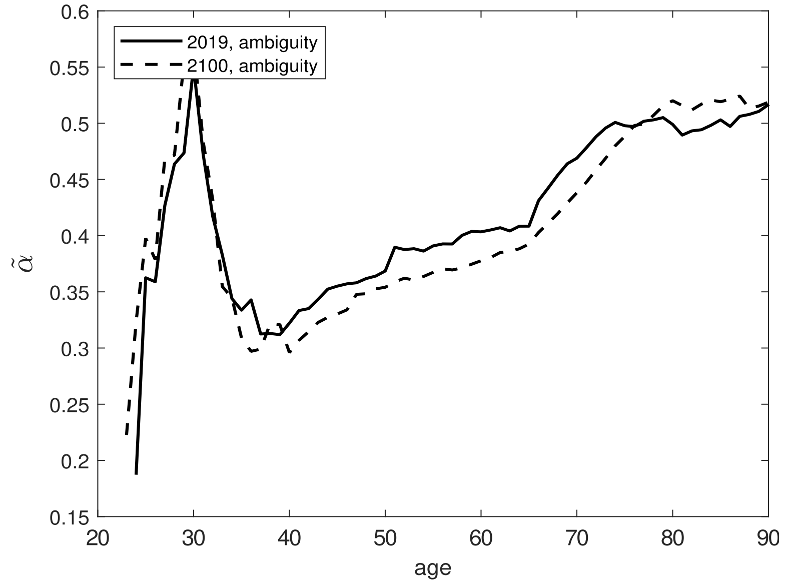
Note: Plots constructed using the calibration and data described in Section 5.1 and Appendix D.2, without any heterogeneity in perceived mortality rates.

D.7 Endogenous equity premium

Figure 30 reproduces Figure 10, allowing for an endogenous response of the equity premium to the increase in life expectancy. Specifically, we assume that the relative supply of risky assets is fixed as in Section 3.5 (equation (54)). We calibrate the relative supply $\bar{S}$ such that the calibrated equity premium used in Section 5 clears the asset market in 2019. We then hold this fixed, and allow the equity premium to vary to ensure that the asset market continues to clear at this relative supply in 2100. To achieve this, we hold the risk-free rate constant, and let the expected risky asset return adjust.

Following the logic laid out for the simple model in Section 3.5, the increase in life expectancy to 2100 implies a modest increase in the equity premium, by 7 basis points. Since this change is small, Figure 30 is very similar to Figure 10. In the analysis in Section 5.3 with a fixed equity premium, the gap between the conditional risky asset shares of those aged 80 and 35 rises from 16.6 p.p. in 2019 to 20.9 p.p. in 2100. With this endogenous response of the equity premium response, the 2100 gap becomes 21.1 p.p..

Figure 30: Model-implied conditional risky asset shares, 2019 and 2100, with endogenous equity premium.



Note: Plots constructed using the calibration and data described in Section 5.1 and Appendix D.2. For the '2100' line, μ is varied such that the relative aggregate demand for risky assets remains at its 2019 level. This involves a rise from the original calibration ($\mu = r^f + 0.04$) to $\mu = r^f + 0.04066$.

Appendix: Chapter 2

A Core Model: Additional Details and Proofs

This Appendix provides additional details on the stylized model of Section 2, as well as all the theoretical proofs.

A.1 Model Details

In this Appendix, we provide the remaining details to our stylized model. In particular, we describe the household problem and formally define the equilibrium.

Household Problem. A representative household owns all businesses in the economy and optimally chooses aggregate consumption, C, such that:

$$C = \sum_j (W_j n_j + \pi_j).$$

where the aggregate price is normalized. $W_j = h(\omega_j)W$ is the firm-specific wage. n_j and π_j are firm j's employment and profit, respectively.

Equilibrium. A stationary equilibrium consists of (i) a value function $v(z)$ and policy functions $n(z)$ and $\omega(z)$, and (ii) a fully on-site wage rate $W \geq 0$, mass of startup attempts $M_e \geq 0$, and a measure of incumbents $\mu(z)$, such that:

- $v(z)$, $n(z)$ and $\omega(z)$ solve the incumbent's problem (62),
- the free entry condition (63) is satisfied with equality if $M_e > 0$,
- the labor and goods markets clear:

$$\begin{aligned} N &= \int n\mu(z)dz \\ Y &= C + \kappa_e M_e + \int g(\omega)(\kappa_n n + \kappa_o)\mu(z)dz, \end{aligned}$$

  where total output, $Y = \int f(\omega)y\mu(z)dz$, is used for household consumption, C, and for entry, non-wage labor and overhead costs.

- and the distribution of firms satisfies: $\mu(z) = h(z)$.

A.2 Proofs

In what follows, we provide all the proofs to our propositions in the main text, assuming that $h(\omega) = g(\omega)$. For simplicity, we first denote $\kappa_n + W = \kappa$.

Proof of Proposition 5. Differentiating $\pi(z)$ w.r.t. ω and n gives the FOCs:

$$\begin{aligned} f'(\omega)zn^{\alpha} - g'(\omega)(\kappa n + \kappa_o) &= 0 \\ f(\omega)z\alpha n^{\alpha-1} - g(\omega)\kappa &= 0 \end{aligned}$$

1. if $\kappa_o = 0$, then combining the two FOCs and rearranging gives ω^* such that:

$$\frac{f'(\omega^*)}{f(\omega^*)}\frac{g(\omega^*)}{g'(\omega^*)} = \alpha,$$

2. if $\kappa_o > 0$, differentiating equations (192) and (193) w.r.t. z and rearranging, we obtain:

$$\{(\frac{f''g' - g''f'}{g'^3})f\alpha(\alpha-1) - [\frac{f'}{g'}(\alpha-1) + \frac{\kappa_o}{zn^{*\alpha}}]^2\}\frac{\partial\omega^*}{\partial z} = \frac{f\alpha\kappa_o}{g'z^2n^{*\alpha}}$$

Since $\kappa_o > 0$ and $g' < 0$, the RHS is negative. Hence $\frac{\partial\omega^*}{\partial z} < 0$ if and only if:

$$(\frac{f''g' - g''f'}{g'^3})f\alpha(\alpha-1) > [\frac{f'}{g'}(\alpha-1) + \frac{\kappa_o}{zn^{*\alpha}}]^2 \tag{187}$$

We can further derive the necessary condition from (187):

$$\frac{f''}{f'} > \frac{g''}{g'}$$

Proof of Proposition 6. Assume two coefficients $\widetilde{f}$ and $\widetilde{g}$ that govern the velocity of productivity loss and cost savings. Specifically, we have: $\frac{\partial f(\omega;\widetilde{f})}{\partial \widetilde{f}} > 0$, $\frac{\partial g(\omega;\widetilde{g})}{\partial \widetilde{g}} < 0$, $\frac{\partial^2 f(\omega;\widetilde{f})}{\partial\omega\partial\widetilde{f}} > 0$ and $\frac{\partial^2 g(\omega;\widetilde{g})}{\partial\omega\partial\widetilde{g}} < 0$ when $\omega \in (0,1]$.

For simplicity, we denote $\frac{\partial f(\omega;\widetilde{f})}{\partial \widetilde{f}}$ as f_2, $\frac{\partial f(\omega;\widetilde{f})}{\partial\omega}$ as f_1, $\frac{\partial^2 f(\omega;\widetilde{f})}{\partial\omega\partial\widetilde{f}}$ as f_{12}, and $\frac{\partial^2 f(\omega;\widetilde{f})}{\partial\omega^2}$ as f_{11}. Similar for $g(\widetilde{g},\omega)$. We can rewrite Equation (192) and (193) as:

$$f_1(\omega^*;\widetilde{f})zn^{*\alpha} - g_1(\omega^*;\widetilde{g})(\kappa n^* + \kappa_o) = 0 \tag{188}$$

$$f(\omega^*;\widetilde{f})z\alpha n^{*\alpha-1} - g(\omega^*;\widetilde{g})\kappa = 0 \tag{189}$$

Proof of $\frac{\partial v_e}{\partial \widetilde{f}} > 0$. By the envelope theorem, we have:

$$\frac{\partial \pi^*}{\partial \widetilde{f}} = f_2(\omega^*; \widetilde{f}) z n^{*\alpha} > 0 \tag{190}$$

Note that:

$$\begin{aligned} v_e &= \int v(z) h(z) dz \\ &= \int \frac{\pi^*(z)}{1-\beta(1-\delta)} h(z) dz \end{aligned}$$

Since $\frac{\partial h(z)}{\partial \widetilde{f}} = 0$ and $h(z) \geq 0$, we have:

$$\frac{\partial v_e}{\partial \widetilde{f}} = \int \frac{\partial \pi^*(z)}{\partial \widetilde{f}} \frac{h(z)}{1-\beta(1-\delta)} dz > 0$$

Proof of $\frac{\partial v_e}{\partial \widetilde{g}} > 0$. By the envelope theorem, we have:

$$\frac{\partial \pi^*}{\partial \widetilde{g}} = -g_2(\omega^*; \widetilde{g})(\kappa n^* + \kappa_o) > 0 \tag{191}$$

Similarly, we have:

$$\frac{\partial v_e}{\partial \widetilde{g}} = \int \frac{\partial \pi^*(z)}{\partial \widetilde{g}} \frac{h(z)}{1-\beta(1-\delta)} dz > 0$$

Interior Solutions. Start with the FOCs:

$$f'(\omega) z n^\alpha - g'(\omega)(\kappa n + \kappa_o) = 0 \tag{192}$$

$$f(\omega) z \alpha n^{\alpha-1} - g(\omega)\kappa = 0 \tag{193}$$

Solve the optimal employment, n^*, from equation (193):

$$n^* = \left(\frac{f z \alpha}{g \kappa}\right)^{\frac{1}{1-\alpha}}$$

Substituting n^* into equation (192), we have

$$\frac{z f'(\omega^*)}{g'(\omega^*)} - \kappa n^{*1-\alpha} - \kappa_0 n^{*-\alpha} = 0 \tag{194}$$

Define the followings for simplicity:

$$\begin{aligned} F(\omega) &= \alpha z f(\omega) \\ G(\omega) &= \kappa g(\omega) \end{aligned}$$

Then we can rewrite equation (194) as:

$$\frac{1}{\alpha}\frac{F'/F}{G'/G} - 1 - \frac{\kappa_0}{\kappa}(\frac{F}{G})^{-\frac{1}{1-\alpha}} = 0$$

Denote $H(\omega) = \frac{1}{\alpha}\frac{F'/F}{G'/G} - 1 - \frac{\kappa_0}{\kappa}(\frac{F}{G})^{-\frac{1}{1-\alpha}}$. By the intermediate value theorem, the sufficient condition for interior solutions is thus:

$$H(0)H(1) < 0$$

B Generalized Model: Additional Details and Results

This Appendix provides a formal definition of equilibrium in the generalized model.

B.1 Equilibrium Definition in Generalized Model

A stationary equilibrium consists of (i) value functions $v(z,k)$, $v^{\omega}(z,k)$, $v^{x}(z,k)$ and policy functions $n(z,k)$, $\omega(z,k)$, $\widetilde{r}(z,k)$, $\widetilde{z}(k)$, $x(z,k)$ and (ii) a fully on-site wage rate $W \geq 0$, a mass of entrants $M \geq 0$ and a measure of incumbents $\overline{\mu}(z,k)$ (with $\mu(z,k)$ denoting the probability distribution), such that:

- $v(z,k)$, $v^{\omega}(z,k)$, $v^{x}(z,k)$, $n(z,k)$, $\omega(z,k)$, $\widetilde{r}(z,k)$, $\widetilde{z}(k)$, $x(z,k)$ solve the incumbent's problem (70);
- the free entry condition (76) is satisfied;
- the goods and labor markets clear (77), (78);
- the distribution of firms satisfies

  $$\overline{\mu}(z',k') = \int\int \Phi(z',k'|z,k)d\overline{\mu}(z,k) + M\mathbb{1}[k' = x(z',0)]H(z'),$$

  where

  $$\Phi(z',k'|z,k) = F(z'|z)\mathbb{1}[k' = x(z,k) + (1-\delta)k(z,k)]\mathbb{1}[\widetilde{z}(k)],$$

  and where $\mathbb{1}[\widetilde{z}(k)]$ is an indicator function equal to 1 when firms decide to remain in operation, $F(z'|z)$ is the transition function for productivity shocks described in (67) and, therefore, where $\Phi(z',k'|z,k)$ denotes the transition from (z,k) to (z',k'). $\widetilde{r}(z,k)$ denotes the decision of first-time work from home.

B.2 Computational Strategy

- Given $\widetilde{f}$, $\widetilde{g}$ and $\widetilde{h}$, guess the equilibrium fully on-site wage W.
- For all pairs (z,k) on the grid, such that $\mu(z,k) > 0$, the optimal choices of $(n_{j,t}, x_{j,t})$ (for onsite firms) and $(n_{j,t}, \omega_{j,t}, x_{j,t})$ (for work-from-home firms) are the solutions to the following problem:

$$
\begin{aligned}
v_j(z_{j,t}, k_{j,t}) &= \max_{n_{j,t}, x_{j,t}} \{\pi_{j,t} + \beta(1-\delta)u_j(z_{j,t}, k_{j,t})\} \\
\pi_{j,t} &= f(\omega_{j,t})y_{j,t} - W(\omega_{j,t})n_{j,t} - g(\omega_{j,t})(\kappa_n n_{j,t} + \mu_\kappa) - x_{j,t} - \zeta(x_{j,t}, k_{j,t}) \\
u_j(z_{j,t}, k_{j,t}) &= \int \max \left[\begin{array}{c} v_j^x(k_{t+1}), \mathbb{E}_t v_j(z_{j,t+1}, k_{t+1}) - \widetilde{\kappa}_o, \\ \mathbb{E}_t v_j^\omega(z_{j,t+1}, k_{j,t+1}) - \widetilde{\kappa}_o - \kappa_j^\omega \end{array} \right] dH_\kappa(\widetilde{\kappa}_o) \\
v_j^x(z_{j,t}, k_{j,t}) &= k_{j,t}(1-\delta_k) - \zeta(-k_{j,t}(1-\delta_k), k_{j,t}) \\
v_j^\omega(z_{j,t}, k_{j,t}) &= \max_{n_{j,t}, \omega_{j,t}, x_{j,t}} \left\{\pi_{j,t} + \beta(1-\delta)u_j^\omega(z_{j,t}, k_{j,t})\right\} \\
u_j^\omega(z_{j,t}, k_{j,t}) &= \int \max \left[v_j^x(k_{t+1}), \mathbb{E}_t v_j^\omega(z_{j,t+1}, k_{t+1}) - \widetilde{\kappa}_o\right] dH_\kappa(\widetilde{\kappa}_o)
\end{aligned}
$$

- Using the free entry condition and the entry function, compute the mass of startup attempts, s_i, and the mass of successful startups, m_i:

$$
\begin{aligned}
\kappa_e &= \frac{m_i}{s_i} \sum_l p_l \int_z \max\left[v(z, 0; \underline{z}_i, \kappa_l^\omega), v^\omega(z, 0; \underline{z}_i, \kappa_l^\omega) - \kappa_l^\omega\right] dH_z(z) \\
m_i &= \Psi_i^\phi s_i^{1-\phi}
\end{aligned}
$$

- Using s_i, m_i, the aggregate resource constraint and consumption FOC to pin down the implied mass M:

$$
\begin{aligned}
Y &= C + S\kappa_e + \sum_i \sum_l \int\int \left[\zeta(x,k) + g(\omega)(\kappa_n n + \mu_\kappa) + \hat{\kappa}_o + \kappa_l^\omega \mathbb{1}_{i,l}(z,k)\right] \mu_{i,l}(z,k) dz dk \\
W &= \upsilon C
\end{aligned}
$$

- Iterate on finding a equilibrium wage such that the following is satisfied:

$$
M = \sum_i m_i
$$

B.3 Solution to the Counterfactual Economy with "Fixed Mass of Firms"

The solution to the counterfactual economy with "no change in the mass of firms" is generated by the following steps:

- The first two steps are the same as in the last subsection.

- Using the free entry condition and the entry function, compute the probability of successful startups.
- Compute the "hypothetical" mass of startups such that the total mass of firms is the same as in the benchmark economy. Use the probability of successful startups obtained from the last step to back out the "hypothetical" mass of startup attempts.
- Using the "hypothetical" mass of startups attempts and successful startups, and the resource constraint to compute the aggregate consumption.
- Iterate on finding an equilibrium fully on-site wage such that the consumption FOC is satisfied.

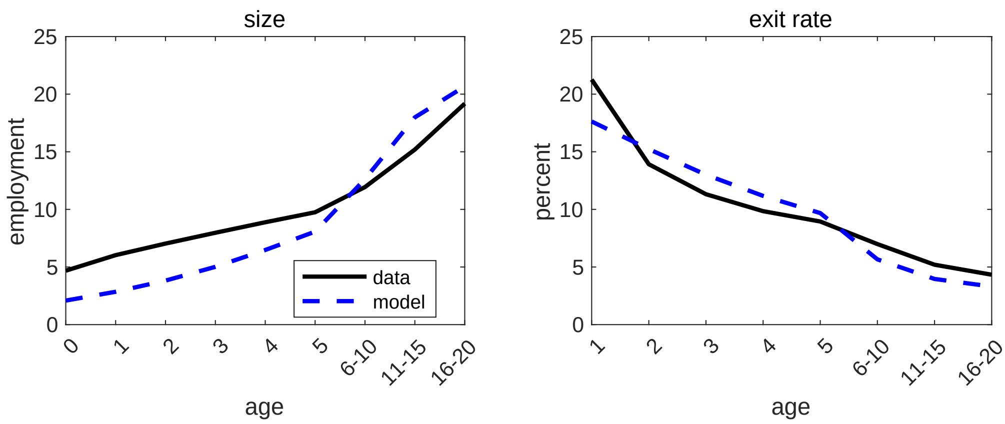
B.4 Robustness: Elasticity of Entry Function

As discussed in the calibration section, the elasticity of the entry function, ϕ, is important for the response of startups to changes in remote work conditions. In this Appendix, we provide a sensitivity analysis with respect to ϕ.

In particular, we consider higher (0.2) and lower (0.1) values of ϕ and re-calibrate both cases to match the same targets as our baseline economy. Table 25 and 28 show the calibrated parameters, respectively. Table 26 to 30 show the results.

While the elasticity of entry matters for the strength of the entry response, it matters less for changes in firm sizes and aggregates. The reason is that a shallower (stronger) entry response is compensated for by a stronger (weaker) change in firm selection. Therefore, aggregate outcomes are effectively identical across the 3 sets of ϕ values.

Figure 31: Life-cycle profiles of size and exit rates: Data and model



Note: The left panel shows average establishment size (employment) by age, while the right panel shows average exit rates by age for the model and the data. The latter is taken from the BED, averaged over the years 2003-2019.

Table 25: Parameter values and targeted moments

Parameter		Value	Target/Source	Data	Model
β	Discount factor	0.96	Interest rate of approx. 4%		
α	Labor elasticity of output	0.65	Labor share in income of approx. 65%		
θ	Returns to scale	0.90	Basu, Fernald (1997) estimate		
δ_k	Capital depreciation	0.08	Cooper, Haltiwanger (2006)		
υ	Disutility of labor	0.018	Normalization, $W = 1$		
κ_e	Entry cost	0.72	Normalization, $s_H + s_L = 1$		
ϕ	Elasticity of entry function	0.10	Sedláček and Sterk (2017)		
κ_L^ω	"Low" remote work setup costs	0	Normalization, minimum remote work setup cost of 0		
$\widetilde{f}$	Speed of remote work productivity losses	-0.151	Average work from home rate, ATUS	4.1%	3.9%
$\widetilde{g}$	Speed of remote work cost savings	0.325	Work from home rates in 100+ over 100- firms, ATUS & ASEC	1.12	1.12
$\widetilde{h}$	Household preferences for remote work	0	Wages unresponsive to remote work	0	0
κ_n	Non-wage labor costs	0.255	Rental expenses - size coefficient, Compustat (see (83))	0.86	0.91
κ_H^ω	"High" remote work setup costs	5.3	Share of large firms conducting remote work, BRS	30%	29%
Ψ_ω	Share of "low" remote work setup costs	0.195	Share of firms conducting remote work, BRS	23%	23%
Ψ_L	Mass of "low" business opportunities	$8.0e-7$	Share of small (<50) businesses, BED	95%	94%
Ψ_H	Mass of "high" business opportunities	$2.4e-8$	Startup success rate, BED	21%	23%
$\underline{z}_H$	Mean "high" productivity	0.131	Average establishment size, BED	15.4	15.0
$\underline{z}_L$	Mean "low" productivity	0.104	Average establishment size of small (<50) businesses, BED	6.8	7.6
ρ	Productivity persistence	0.723	Establishment size life-cycle profile, BED	see Figure 31	
σ_z	Productivity shock dispersion	0.208	Establishment size life-cycle profile, BED	see Figure 31	
ζ_0	Capital adjustment costs, fixed	0.001	Establishment size life-cycle profile, BED	see Figure 31	
ζ_1	Capital adjustment costs, variable	0.61	Establishment size life-cycle profile, BED	see Figure 31	
δ	Exogenous exit rate	0	Establishment exit life-cycle profile, BED	see Figure 31	
μ_κ	Overhead costs, mean	0.795	Establishment exit life-cycle profile, BED	see Figure 31	
σ_κ	Overhead costs, dispersion	2.49	Establishment exit life-cycle profile, BED	see Figure 31	

Note: The table presents values for all model parameters. The first three columns describe each parameter and its parameterized value, while the last two columns indicate the most relevant target or data source.

Table 26: Model Results: Remote work and business entry and exit ($\phi = 0.1$)

	Rate		Size	
	Entry	Exit	Entrants	Exiters
Data	+28%	+15%	−24%	−22%
Model	+19%	+19%	−19%	−25%

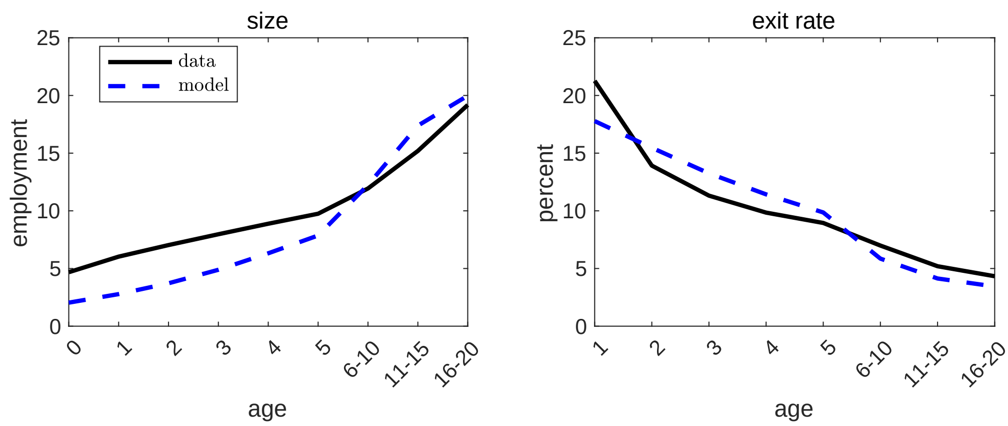
Note: The table shows changes in entry and exit rates (first two columns) and changes in average of entrants and of exiting businesses (last two columns). The top row shows changes in the BED data, while the bottom row shows the model-predicted changes.

Table 27: Impact of increased remote work: Changes in aggregates ($\phi = 0.1$)

	Output (Y)		Consumption (C)			Welfare ($\mathcal{W}$)	
Overall	3.2		4.7			0.2	
Components	Ω	$\overline{y}$	Y	I	Costs	C	N
	28.9	−25.6	4.8	−1.0	0.9	0.3	−0.1

Note: The first row of the table shows log-changes in aggregate Output (Y), Consumption (C) and Welfare ($\mathcal{W}$). The second row then split the overall changes into the contributions of the various components. All values are reported in percent changes from the pre-pandemic baseline.

Figure 32: Life-cycle profiles of size and exit rates: Data and model



Note: The left panel shows average establishment size (employment) by age, while the right panel shows average exit rates by age for the model and the data. The latter is taken from the BED, averaged over the years 2003-2019.

Table 28: Parameter values and targeted moments

	Parameter	Value	Target/Source	Data	Model
β	Discount factor	0.96	Interest rate of approx. 4%		
α	Labor elasticity of output	0.65	Labor share in income of approx. 65%		
θ	Returns to scale	0.90	Basu, Fernald (1997) estimate		
δ_k	Capital depreciation	0.08	Cooper, Haltiwanger (2006)		
υ	Disutility of labor	0.022	Normalization, $W = 1$		
κ_e	Entry cost	0.63	Normalization, $s_H + s_L = 1$		
ϕ	Elasticity of entry function	0.20	Sedláček and Sterk (2017)		
κ_L^ω	"Low" remote work setup costs	0	Normalization, minimum remote work setup cost of 0		
$\widetilde{f}$	Speed of remote work productivity losses	-0.151	Average work from home rate, ATUS	4.1%	3.9%
$\widetilde{g}$	Speed of remote work cost savings	0.325	Work from home rates in 100+ over 100- firms, ATUS & ASEC	1.12	1.12
$\widetilde{h}$	Household preferences for remote work	0	Wages unresponsive to remote work	0	0
κ_n	Non-wage labor costs	0.255	Rental expenses - size coefficient, Compustat (see (83))	0.86	0.91
κ_H^ω	"High" remote work setup costs	5.3	Share of large firms conducting remote work, BRS	30%	29%
Ψ_ω	Share of "low" remote work setup costs	0.195	Share of firms conducting remote work, BRS	23%	23%
Ψ_L	Mass of "low" business opportunities	$3.8e-4$	Share of small (< 50) businesses, BED	95%	94%
Ψ_H	Mass of "high" business opportunities	$3.8e-5$	Startup success rate, BED	21%	21%
$\underline{z}_H$	Mean "high" productivity	0.131	Average establishment size, BED	15.4	14.3
$\underline{z}_L$	Mean "low" productivity	0.104	Average establishment size of small (< 50) businesses, BED	6.8	7.4
ρ	Productivity persistence	0.723	Establishment size life-cycle profile, BED	see Figure 32	
σ_z	Productivity shock dispersion	0.208	Establishment size life-cycle profile, BED	see Figure 32	
ζ_0	Capital adjustment costs, fixed	0.001	Establishment size life-cycle profile, BED	see Figure 32	
ζ_1	Capital adjustment costs, variable	0.61	Establishment size life-cycle profile, BED	see Figure 32	
δ	Exogenous exit rate	0	Establishment exit life-cycle profile, BED	see Figure 32	
μ_κ	Overhead costs, mean	0.795	Establishment exit life-cycle profile, BED	see Figure 32	
σ_κ	Overhead costs, dispersion	2.49	Establishment exit life-cycle profile, BED	see Figure 32	

Note: The table presents values for all model parameters. The first three columns describe each parameter and its parameterized value, while the last two columns indicate the most relevant target or data source.

Table 29: Model Results: Remote work and business entry and exit ($\phi = 0.2$)

	Rate		Size	
	Entry	Exit	Entrants	Exiters
Data	+28%	+15%	−24%	−22%
Model	+11%	+11%	−13%	−19%

Note: The table shows changes in entry and exit rates (first two columns) and changes in average of entrants and of exiting businesses (last two columns). The top row shows changes in the BED data, while the bottom row shows the model-predicted changes.

Table 30: Impact of increased remote work: Changes in aggregates ($\phi = 0.2$)

	Output (Y)		Consumption (C)			Welfare ($\mathcal{W}$)	
Overall	2.6		4.1			0.2	
Components	Ω	$\overline{y}$	Y	I	Costs	C	N
	20.7	−18.1	3.9	−0.8	1.1	0.3	−0.1

Note: The first row of the table shows log-changes in aggregate Output (Y), Consumption (C) and Welfare ($\mathcal{W}$). The second row then split the overall changes into the contributions of the various components. All values are reported in percent changes from the pre-pandemic baseline.

C Empirical Analysis: Additional Exercises and Robustness

In this Appendix, we consider various robustness checks to some of our empirical results as well as a set of additional estimates.

C.1 Heterogeneity of Remote Work Across Sectors

According to the Business Response Survey, the fraction of establishments with some employees working remotely was 23.3% in February 2020. Turning to the intensive margin of remote work, the American Time Use Survey suggests that the fraction of hours worked remotely in 2019 was "only" about 6.7%.

Both these averages, however, hide large amounts of heterogeneity across sectors. Intuitively, remote work – both at the extensive and intensive margins – is most common in the service sectors. For instance, the information sector is characterized by almost 59% of establishments reporting some employees working remotely with an average 8.8% of hours worked from home. Professional and business services, as well as financial services have similarly high rates of remote work. In contrast, accommodation and food services have the lowest remote work rates with only 2.1% of establishments reporting some employees working remotely and an average 1.3% of hours worked from home. Construction and retail trade have similarly low levels of remote work.

C.2 Work from Home and Business Dynamism

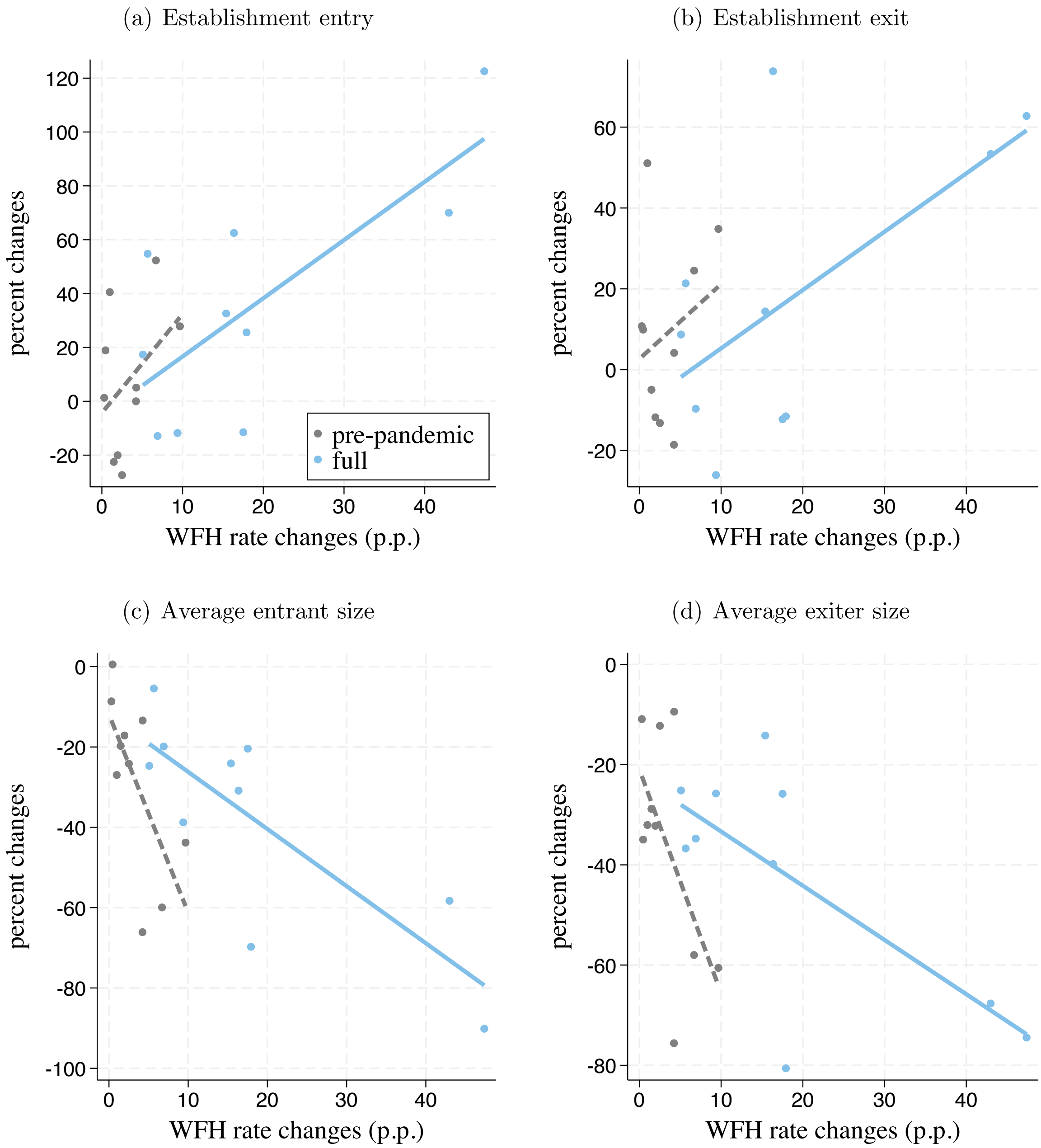
We now turn to the link between work from home rates and business entry and exit.

Raw data. Figure 33 shows how changes in remote work rates relate to business entry, exit, the average size of entering and exiting establishments. Specifically, the horizontal axis shows percentage point changes in industry-level work from home rates, while the vertical axis depicts the corresponding percent changes in the numbers of entrants and exiters, and their respective sizes. In all these cases, we consider separately the full sample (2003-2022) and the pre-pandemic period (2003-2019).[40]

The figure shows that – consistent with our core theoretical model – increases in work from home rates are clearly associated with strong increases in firm entry and exit, coupled with declines in the size of entering and exiting businesses. Note that these patterns are

[40] We compute changes as the difference in the respective values at the end of the pre-pandemic period (average of 2018Q1-2019Q4) or the full sample (average of 2021Q1-2022Q4) relative to the start of our sample (average of 2003Q1-2004Q4).

Figure 33: Work from home and business dynamism: Changes across industries



Note: The figure depicts super-sector changes in work from home (WFH) rates on the horizontal axis (in percentage points) and (percent) changes in the number of entrants (Panel a), exiters (Panel b), average entrant size (Panel c) and average exiter size (Panel d). Work from home rates are estimated from ATUS as described in the main text. Business entry, exit and sizes are taken from the BED. All panels show data for the pre-pandemic sample (2003-2019) and the full sample (2003-2022).

not a pandemic-only phenomenon. In fact, the relationship is somewhat weaker during the pandemic which saw unprecedented increases in work from home rates.

Table 31: Working from home and business dynamism: Regression results

	Entry	Exit	Entrant Size	Exiter Size
A: Pre-pandemic sample (2003Q1-2019Q4)				
Work from home rate, β	1.414^{***}	1.214^{***}	-1.315^{***}	-1.235^{***}
	(0.218)	(0.240)	(0.191)	(0.180)
R-squared	0.502	0.405	0.420	0.566
# observations	590	590	590	590
B: Full sample (2003Q1-2022Q4)				
Work from home rate, β	1.400^{***}	0.905^{***}	-0.712^{***}	-0.408^{***}
	(0.117)	(0.120)	(0.106)	(0.085)
R-squared	0.705	0.547	0.486	0.581
# observations	710	700	710	700

Note: The table reports results from estimating (195). Panel A reports estimates using the pre-pandemic period (2003Q1-2019Q4) only, while Panel B report results for our entire sample period (2003Q1-2022Q4). All specifications include industry and time fixed effects as well as lagged values of the dependent variable and industry-level real output growth rates. Standard errors are reported in brackets – all estimates are significant at the 1 percent level.

Estimation. To test the above relationships more formally, we estimate the following panel regressions:

$$y_{i,t} = \delta_i + \delta_t + \beta\overline{\omega}_{i,t} + \Gamma X_{i,t} + \epsilon_{i,t}, \tag{195}$$

where $y_{i,t}$ represents, respectively, (the logs of) business entry, exit, average entrant's or exiter's size in industry i and period t, δ_i are industry fixed effects, δ_t are time fixed effects, $X_{i,t}$ is a set of control variables and $\overline{\omega}^L_{i,t} = 1/(L+1)\sum_{l=0}^{L}\omega_{i,t-l}$ are time-varying moving averages of work from home rates. Coefficient β is the primary object of interest as it provides a concise summary of the potentially dynamic (lagged) effects of working from home rates on business dynamism.[41]

In estimating β, we control for a range of variables. First, $X_{i,t}$ includes lags of our (average) work from home measure, $\overline{\omega}$. Second, in addition to controlling for industry differences through fixed effects, δ_i, and aggregate trends through time fixed effects, δ_t, we also include industry-specific real output growth rates, $g_{i,t}$, taken from the Bureau of Economic Analysis. Finally, as before, we consider two sample periods for our specifications: "pre-pandemic" sample and "full" sample.

Table 31 shows that even after controlling for a range of other factors, changes in remote work rates are strongly related to changes in business dynamism. Moreover, the direction of these relationships remains the same as in the raw data. In particular, higher remote work rates are related to more business entry and exit, but smaller sizes of entrants and exiting businesses.

[41] In our baseline specification we use $L = 4$. The Appendix provides robustness exercises with respect to L.

C.3 Robustness: Working Outside the Workplace

In this Appendix, we use work-outside-workplace rate to replace work from home rate in the empirical analysis. The construction of work-outside-workplace is similar to that of work from home rate, defined in equation (82), in that we count a day as work outside workplace if the individual spent in total at least 6 hours working at home or other places except their workplace.

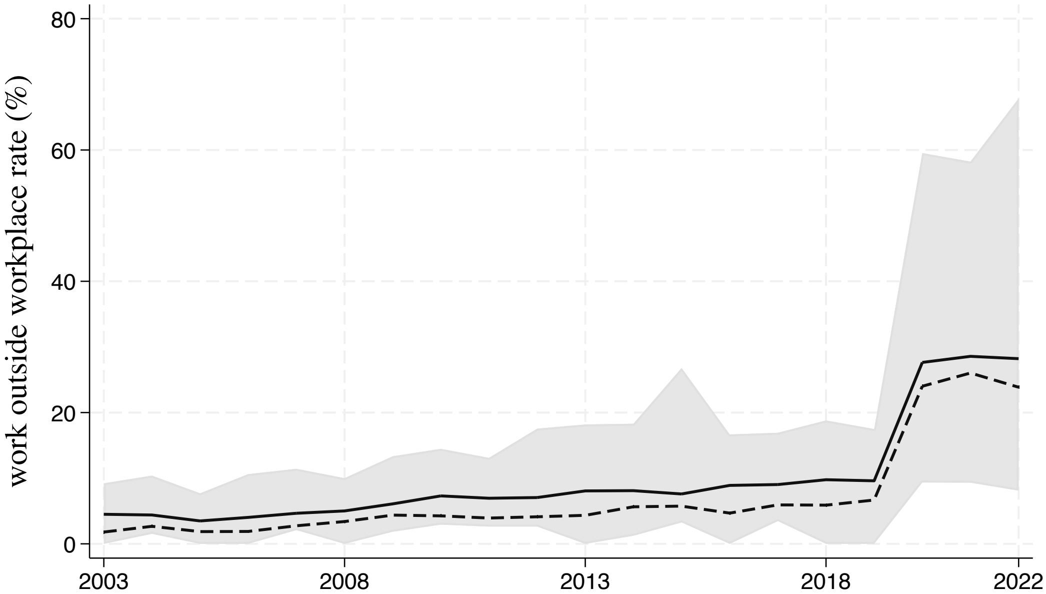
Figure 34 shows the evolution of working-outside-workplace rate from 2003 to 2022. In addition, Figure 35 plots how changes in remote work rates are connected to changes in establishment entry and exit across industries. Finally, Table 32 shows the associated panel regression results. All these are very similar to the outcomes presented in the main text which are based on work from home definitions.

Table 32: Working outside workplace and business dynamism: Regression results

	Entry	Exit	Entrant Size	Exiter Size
A: Pre-pandemic sample (2003Q1-2019Q4)				
Work from home rate, β	0.832***	0.875***	−0.519***	−0.674***
	(0.187)	(0.202)	(0.167)	(0.156)
R-squared	0.465	0.384	0.351	0.526
# observations	590	590	590	590
B: Full sample (2003Q1-2022Q4)				
Work from home rate, β	1.299***	0.924***	−0.723***	−0.393***
	(0.116)	(0.115)	(0.104)	(0.082)
R-squared	0.685	0.538	0.463	0.569
# observations	710	700	710	700

Note: The table reports results from estimating (195). Panel A reports estimates using the pre-pandemic period (2003Q1-2019Q4) only, while Panel B report results for our entire sample period (2003Q1-2022Q4). All specifications include industry and time fixed effects as well as lagged values of the dependent variable and industry-level real output growth rates. Standard errors are reported in brackets – all estimates are significant at the 1 percent level.

Figure 34: Work outside workplace rate: Changes over time



Note: The figure shows work outside workplace rates over time for the aggregate economy (solid black line) and the range of values across industries (shaded area). The work from home rates over time for the aggregate economy (dashed black line) is added for comparison.

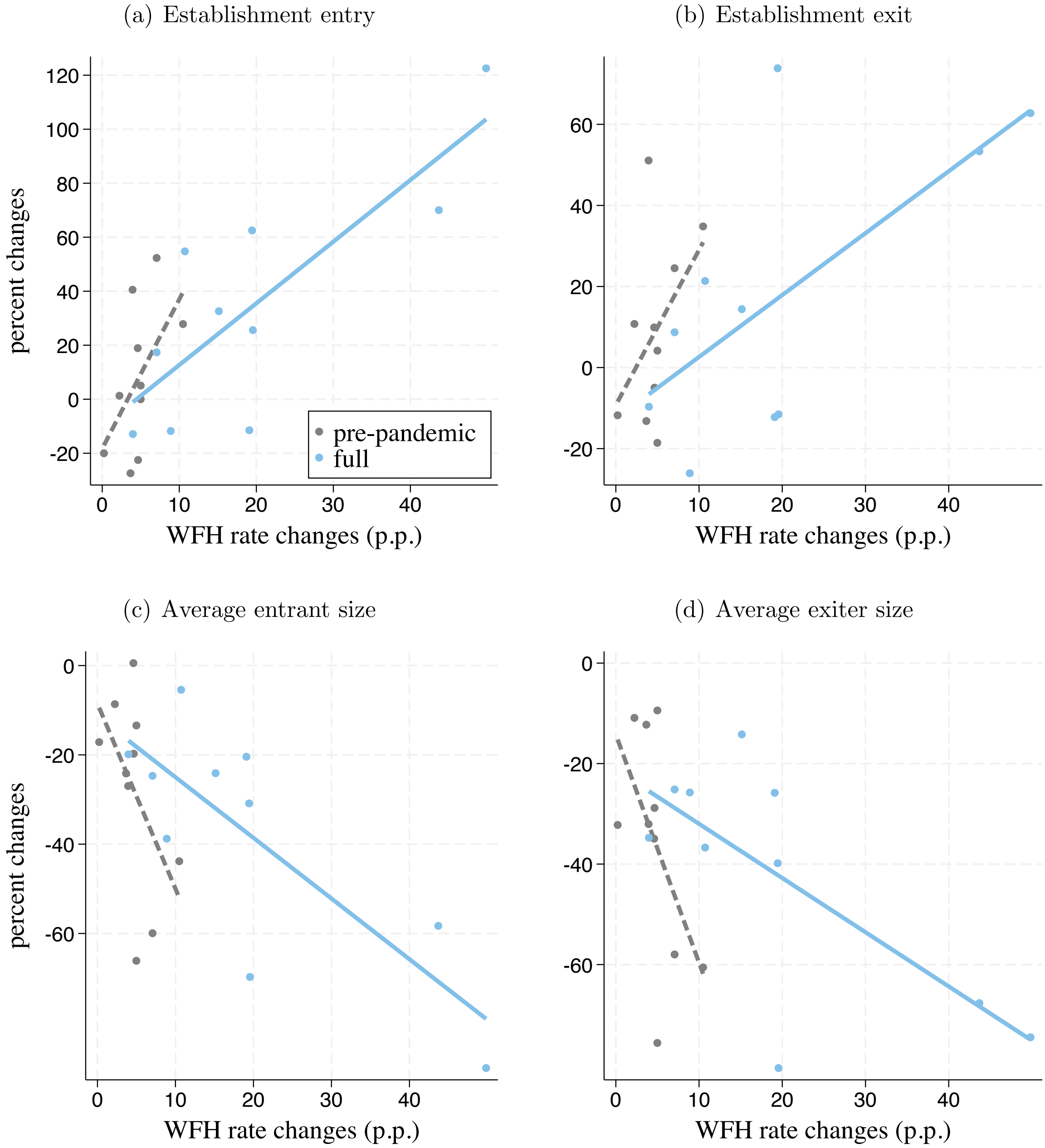
C.4 Robustness: Using Average Size

In the main text, we use average entrant size as a measurement of employment. Here we provide the results using average size computed from QCEW. Figure 36 shows how the change in remote work is associated with changes in average establishment size. Table 33 shows the results of panel regression.

Table 33: Working from home and establishment size

	Average Size
A: Pre-pandemic sample (2003Q1-2019Q4)	
Work from home rate, β	-0.604^{***}
	(0.236)
R-squared	0.115
# observations	590
B: Full sample (2003Q1-2022Q4)	
Work from home rate, β	-0.605^{***}
	(0.110)
R-squared	0.256
# observations	710

Figure 35: Work outside workplace and business dynamism: Changes across industries

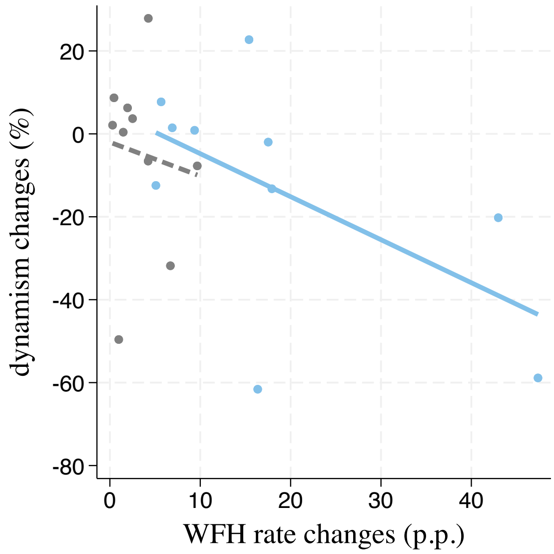


Note: The figure depicts super-sector changes in work outside workplace rates on the horizontal axis (in p.p.) and (percent) changes in the number of entrants (Panel a), exiters (Panel b), average entrant size (Panel c) and average exiter size (Panel d). All panels show data for the pre-pandemic sample (2003-2019) and the full sample (2003-2022).

C.5 Robustness: 2-digit Sectors

We use annual establishment age data at 2-digit level from BED, where establishment entry is reflected in the number of establishments of age less than one year. The average entrants size can be computed using the corresponding employment. The information

Figure 36: Work from home and establishment size: Changes across industries



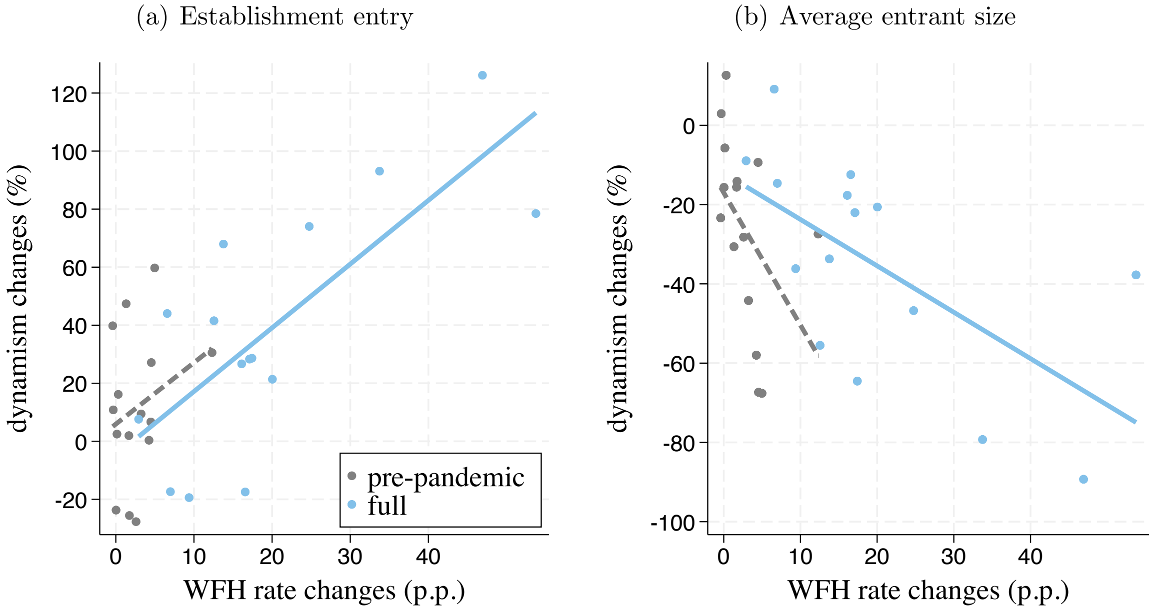
Note: The figure depicts super-sector changes in work from home rates on the horizontal axis (in p.p.) and (percent) changes in the average establishment size. Data is for the pre-pandemic sample (2003-2019) and the full sample (2003-2022).

on establishment exit cannot be deduced from the age data as it would be mixed with temporary closings and reopening. We dropped "Agriculture, forestry, fishing, and hunting", "Mining, quarrying, and oil and gas extraction" and "Management of companies and enterprises", due to limited observations in ATUS. Besides, "Finance and insurance" sector is excluded, consistent with the previous analysis at the super sector level. Figure 37 shows the linkage between work from home rates and business entry. Table 34 shows the results of fixed effect regression, where the average WFH rate is constructed with two lags, i.e., average of the current and the previous two years' WFH rate.

Table 34: Working from home and business entry: 2-digit sectors

	Entry	Entrant Size
A: Pre-pandemic sample (2003Q1-2019Q4)		
Work from home rate, β	1.118***	−1.169***
	(0.415)	(0.328)
R-squared	0.524	0.361
# observations	225	225
B: Full sample (2003Q1-2022Q4)		
Work from home rate, β	2.079***	−1.147***
	(0.199)	(0.156)
R-squared	0.718	0.521
# observations	270	270

Figure 37: Work from home and business entry: Changes across 2-digit sectors



Note: The figure depicts 2-digit-sector changes in work from home rates on the horizontal axis (in p.p.) and (percent) changes in the number of entrants (Panel a) and average entrant size (Panel b). All panels show data for the pre-pandemic sample (2003-2019) and the full sample (2003-2022).

C.6 Robustness: Openings and Closings

As discussed in the main text, BED establishment openings include both births and re-openings, while establishment closings include both deaths and temporary closings. Here we use quarterly establishment openings and closings at the super sector level, consistent with the analysis in the main text. Table 35 reports the results. In Table 36, we further investigate the 2-digit scenario.

C.7 Robustness: Different Lag Lengths

In the main text, we use the current quarter and the last year's WFH rates to construct the regressor. To further validate the lagged impacts of working from home on business entry and exit, we consider $L = 2$ and $L = 6$ in constructing the average WFH rate. Table 37 and 38 report the results.

Table 35: Working from home, establishment openings and closings (super sectors)

	Openings	Closings
A: Pre-pandemic sample (2003Q1-2019Q4)		
Work from home rate, β	0.997***	0.854***
	(0.176)	(0.177)
R-squared	0.440	0.466
# observations	590	590
B: Full sample (2003Q1-2022Q4)		
Work from home rate, β	0.951***	0.682***
	(0.099)	(0.098)
R-squared	0.701	0.651
# observations	710	710

Note: The table reports results from estimating (195). Panel A reports estimates using the pre-pandemic period (2003Q1-2019Q4) only, while Panel B report results for our entire sample period (2003Q1-2022Q4). All specifications include industry and time fixed effects as well as lagged values of the dependent variable and industry-level real output growth rates. Standard errors are reported in brackets – all estimates are significant at the 1 percent level.

Table 36: Working from home, establishment openings and closings (2-digit sectors)

	Openings	Closings
A: Pre-pandemic sample (2003Q1-2019Q4)		
Work from home rate, β	0.428***	0.333***
	(0.149)	(0.153)
R-squared	0.348	0.350
# observations	756	756
B: Full sample (2003Q1-2022Q4)		
Work from home rate, β	1.093***	0.809***
	(0.072)	(0.074)
R-squared	0.689	0.621
# observations	923	923

Note: The table reports results from estimating (195). Panel A reports estimates using the pre-pandemic period (2003Q1-2019Q4) only, while Panel B report results for our entire sample period (2003Q1-2022Q4). All specifications include industry and time fixed effects as well as lagged values of the dependent variable and industry-level real output growth rates. Standard errors are reported in brackets – all estimates are significant at the 1 percent level.

Table 37: Working from home and business dynamism ($L = 2$)

	Entry	Exit	Wages	Entrant Size
A: Pre-pandemic sample (2003Q1-2019Q4)				
Work from home rate, β	0.756***	0.772***	0.410***	−0.820***
	(0.165)	(0.181)	(0.069)	(0.141)
R-squared	0.473	0.381	0.680	0.418
# observations	590	590	590	590
B: Full sample (2003Q1-2022Q4)				
Work from home rate, β	0.991***	0.591***	0.311***	−0.579***
	(0.105)	(0.108)	(0.044)	(0.094)
R-squared	0.692	0.532	0.719	0.479
# observations	710	700	710	710

Note: The table reports results from estimating (195). Panel A reports estimates using the pre-pandemic period (2003Q1-2019Q4) only, while Panel B report results for our entire sample period (2003Q1-2022Q4). All specifications include industry and time fixed effects as well as lagged values of the dependent variable and industry-level real output growth rates. Standard errors are reported in brackets – all estimates are significant at the 1 percent level.

Table 38: Working from home and business dynamism ($L = 6$)

	Entry	Exit	Wages	Entrant Size
A: Pre-pandemic sample (2003Q1-2019Q4)				
Work from home rate, β	1.606***	1.425***	0.795***	−1.695***
	(0.257)	(0.289)	(0.108)	(0.233)
R-squared	0.522	0.446	0.724	0.403
# observations	550	550	550	550
B: Full sample (2003Q1-2022Q4)				
Work from home rate, β	1.532***	1.163***	0.394***	−0.766***
	(0.128)	(0.133)	(0.055)	(0.119)
R-squared	0.722	0.561	0.725	0.473
# observations	670	660	670	670

Note: The table reports results from estimating (195). Panel A reports estimates using the pre-pandemic period (2003Q1-2019Q4) only, while Panel B report results for our entire sample period (2003Q1-2022Q4). All specifications include industry and time fixed effects as well as lagged values of the dependent variable and industry-level real output growth rates. Standard errors are reported in brackets – all estimates are significant at the 1 percent level.

C.8 Robustness: Remote work exposure

In this Appendix, we provide additional robustness with regard to our novel identification strategy in Section 6.

Sectors where remote work is more feasible. First, we test whether the exposed firms experienced an even stronger decline in rental costs if they operated in industries where remote work is particularly feasible. Towards this end, we make use of a classification of occupations (and in turn industries) in terms of their remote work feasibility from Dingel and Neiman (2020):

$$\ln\left(\frac{rent_{j,t}}{n_{j,t}}\right) = \alpha + \beta_E \mathbb{1}(E)_j + \beta_T \mathbb{1}(T)_{j,t} + \beta_H \mathbb{1}_j(H) + \mathbb{1}(T)_{j,t}\left(\beta_{E,T}\mathbb{1}(E)_j + \beta_{E,H,T}\mathbb{1}(H)_j\mathbb{1}(E)_j\right) + \mathbf{X}'\theta + \epsilon_{j,t}, \tag{196}$$

where $\mathbb{1}(H)_j$ is an indicator function equal to one if firm j operates in one of the 5 industries with the highest feasibility of remote work. Our model predicts that $\beta_{0,H} < 0$ as exposed firms in "high-feasibility" industries are particularly positioned to take advantage of the cost-saving nature of remote work.

Table 39 shows the results. Indeed, exposed firms operating in sectors where remote work is particularly feasible experienced stronger declines in rental costs. That said, direct exposure itself remains to lead to rental cost declines in a statistically significant manner.

Table 39: Estimation results: Rental costs, $rent/n$

	(1)	(2)
Exposed firms, $\beta_{E,T}$	-0.38^{***}	-0.47^{***}
	(0.05)	(0.07)
Exposed firms in high-industries, $\beta_{E,H,T}$	-0.80^{***}	-0.50^{**}
	(0.09)	(0.21)
Fixed effects		✓
R-squared	0.02	0.43
Observations	22,903	22,903

Note: The table presents the key coefficients of interest from regression (196). Fixed effects include sector, city and year fixed effects. Standard errors are clustered at the firm level and reported in brackets.

Event study design. We use an event-study design to estimate the impact of the remote work revolution on firms' rental costs:

$$\ln(rent_{j,t}/n_{j,t}) = \alpha + \beta_{0Y}\mathbb{1}_{0Y} + \beta_{5Y}\mathbb{1}_{5Y} + \sum_{k=-3}^{3}\mathbb{1}_{t=2019+k}\left(\mathbb{1}_{0Y}\beta_{0,k} + \mathbb{1}_{5Y}\beta_{5,k}\right) + \mathbf{X}'\theta + \epsilon_{j,t}, \tag{197}$$

where $rent_{j,t}$ again represent rental expenses of firm j in year t, $\mathbb{1}_{0Y}$ ($\mathbb{1}_{5Y}$) is an indicator function equal to one if firm j in 2019 had no rental commitments in future years (had rental commitments five years into the future) and zero otherwise and where $\mathbf{X}_{j,t}$ include sector, city and year fixed effects.

The key coefficients of interest are $\beta_{0,k}$ and $\beta_{5,k}$. These highlight the difference in per-worker rental costs between "exposed" firms which can flexibly adjust their rental costs in face of the remote work revolution ($\beta_{0,k}$) and "non-exposed" firms, i.e. those which cannot flexibly adjust their rental commitments ($\beta_{5,k}$). In addition, we also consider a specification in which the above indicator functions are interacted with sectors in which remote work is particularly feasible based on the occupations which dominate them (see Dingel and Neiman, 2020):

$$\ln(rent_{j,t}/n_{j,t}) = \alpha + \beta_L \mathbb{1}_L + \beta_H \mathbb{1}_H + \beta_{1Y}\mathbb{1}_{1Y} + \beta_{5Y}\mathbb{1}_{5Y} \tag{198}$$

$$+ \sum_{k=-3}^{3} \mathbb{1}_{t=2019+k}\mathbb{1}_L \left(\mathbb{1}_{0Y}\beta^L_{0,k} + \mathbb{1}_{5Y}\beta^L_{5,k}\right) \tag{199}$$

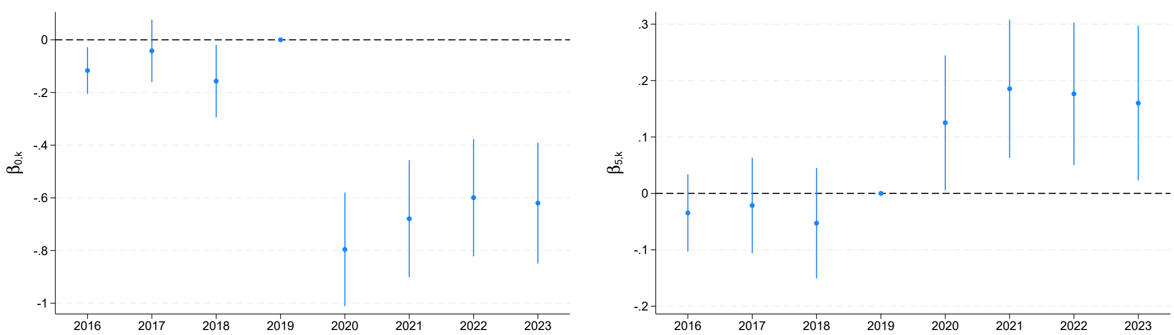
$$+ \sum_{k=-3}^{3} \mathbb{1}_{t=2019+k}\mathbb{1}_H \left(\mathbb{1}_{0Y}\beta^H_{0,k} + \mathbb{1}_{5Y}\beta^H_{5,k}\right) + \mathbf{X}'\theta + \epsilon_{j,t}, \tag{200}$$

where $\mathbb{1}_L$ ($\mathbb{1}_H$) is an indicator which is equal to one for the 5 industries with the lowest (highest) feasibility of remote work.

Figure XYZ presents the results showing that exposed firms experienced a strong drop in rental costs during the pandemic. Moreover, this was further exacerbated for firms in sectors with high feasibility of remote work.

Figure 38: Event study: Rental costs, $rent/n$

(a) Without rental commitments, $\beta_{0,k}$ (b) With 5-year rental commitments, $\beta_{5,k}$



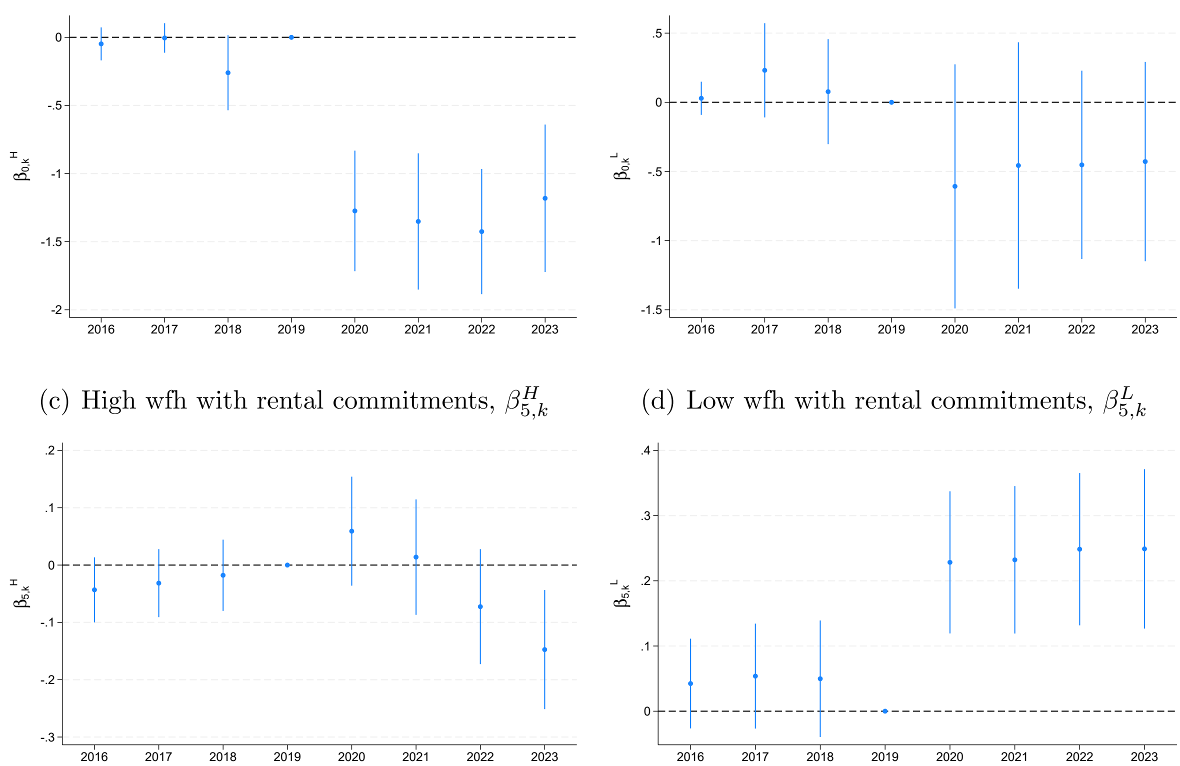
Note: This figure shows the regression results for equation (197).

Firm fixed effects. Next, we consider a specification of our main regression with the addition of firm fixed effects. Table 40 shows that are main results are largely robust to

Figure 39: Event study: Rental costs, $rent/n$, in sectors with high and low wfh-feasibility

(a) High wfh without rental commitments, $\beta^H_{0,k}$ (b) Low wfh without rental commitments, $\beta^L_{0,k}$

(c) High wfh with rental commitments, $\beta^H_{5,k}$ (d) Low wfh with rental commitments, $\beta^L_{5,k}$



Note: This figure shows the regression results for equation (198).

this alternative specification.

Table 40: Estimation results: Firm fixed effects

	rental costs		fixed costs		variable costs		productivity	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Exposure, $\beta_{E,T}$	-0.47^{***}	-0.66^{***}	-0.19^{**}	-0.03	-0.10	-0.05	-0.07^{*}	-0.06^{**}
	(0.05)	(0.06)	(0.08)	(0.03)	(0.08)	(0.03)	(0.04)	(0.03)
Fixed effects		✓		✓		✓		✓
R-squared	0.02	0.05	0.06	0.05	0.0002	0.02	0.01	0.01
Observations	23,003	23,003	23,016	23,016	25,679	25,679	25,139	25,139

Note: The table presents the coefficients of interest on exposed firms in periods 2020-2023 from regressions (85), $\beta_{E,T}$. Fixed effects include sector, city and year fixed effects. Standard errors (reported in brackets) are clustered at the firm level.

Placebo estimation. Table 41 shows results using our estimation in the main text, but estimating it on firms in the period between 2010-2017 (with exposure defined as no commitments in 2014). In contrast to our results in the main text, the placebo treatment does not uncover any significantly negative coefficients.

Table 41: Estimation results: 2010-2017 Placebo

	rental costs		fixed costs		variable costs		productivity	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Exposure, $\beta_{E,T}$	0.05	0.05	-0.06	-0.11	0.17^{*}	-0.01	0.02	-0.002
	(0.06)	(0.07)	(0.10)	(0.07)	(0.10)	(0.05)	(0.05)	(0.04)
Fixed effects		✓		✓		✓		✓
R-squared	0.0001	0.44	0.14	0.58	0.01	0.52	0.03	0.43
Observations	22,932	22,932	22,952	22,952	25,882	25,882	25,302	25,302

Note: The table presents the coefficients of interest on exposed firms in periods 2020-2023 from regressions (85), $\beta_{E,T}$. Fixed effects include sector, city and year fixed effects. Standard errors (reported in brackets) are clustered at the firm level.

Descriptive statistics. Table 42 divides the firms in the sample into four categories: firms reporting rental expenses only, rental commitments only, both rental expenses and commitments, and neither of them. We further shows in each category, the fraction of firms in the entire sample, their average age, median size, median total assets, and average buildings over assets ratio.

Table 42: Descriptive Statistics: Raw Data (2016-2023)

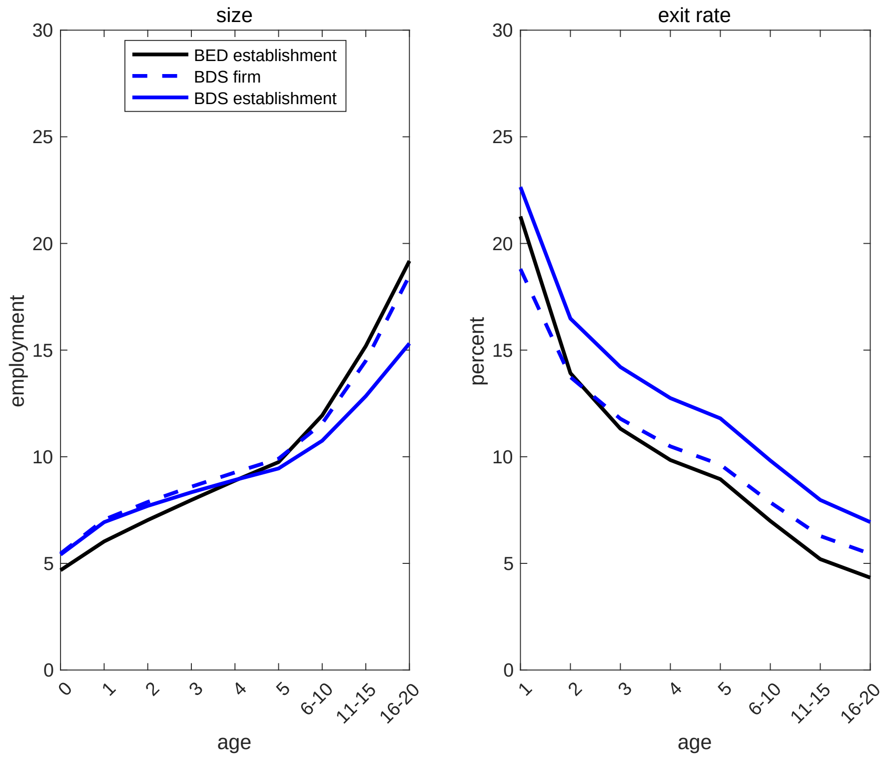
Firms reporting	fraction	age	size	total assets	buildings/assets
Rental expenses	0.52	18.1	1.02	945.6	0.050
Rental commitments	0.40	18.3	0.99	716.4	0.052
Both	0.37	18.5	1.14	771.1	0.052
Neither	0.46	10.9	0.21	1131.9	0.047

Note: Size (Compustat item emp) is in thousands. Total assets (Compustat item at) is in millions. Buildings represent Compustat item fatb, i.e., Property, Plant and Equipment - Buildings at Cost. Employment and total assets are reported in median values. Age and buildings over assets ratio are reported in mean values.

C.9 Comparison between BED and BDS Data

Although we use BED data at the establishment level for calibration, we provide a comparison between BED and BDS data here. From Figure 40, life-cycle profiles of size and exit rates of BED establishments are close to those of BDS firms.

Figure 40: Life-cycle profiles of size and exit rates: BED and BDS



Note: The left panel shows average establishment size from BED and firm/establishment size from BDS by age, while the right panel shows average exit rates by age.

C.10 Aligning Compustat with the BED

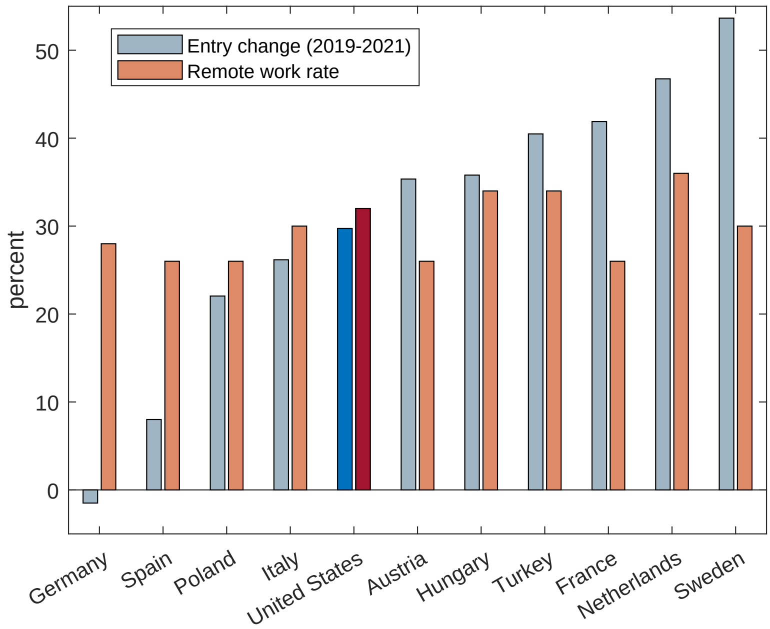
Compustat sample consists only of publicly listed companies that differ significantly from the universe of firms in the U.S. (i.e. the BED data and our model, which targets exit and size profiles in the latter). To address this issue, we follow Ignaszak and Sedláček (2023) and construct a set of firm-specific weights based on the firm size (employment) distribution in Compustat.

Aligning model-simulated data with Compustat information. Formally, we pool all observations across all years and calculate the empirical distribution of firm size in Compustat. Let f_e denote the frequency of firm size e in Compustat. Given that our model is parameterized to business dynamics from the BED, we need to realign the model distribution to that of Compustat whenever comparing model moments to those which have counterparts in Compustat. Towards this end, in any "Compustat-related" regression in the model we weigh the model-simulated data with the respective empirical weights from Compustat, f_e.

C.11 Cross-country comparison

Figure 41 shows a comparison between remote work rates and firm entry across countries. The figure shows that while the increase in remote work rates was similar across countries, entry changes were not. Our model suggests that this is indicative of cross-country differences in the welfare effects of the remote work revolution.

Figure 41: Business entry and remote work across countries



Note: The figure shows changes in firm entry between 2019 and 2021 and remote work rates in 2021 across countries. While the entry data is taken from Eurostat, remote work rates are taken from Aksoy et al. (2022).

Appendix: Chapter 3

A Model Details

A.1 Proof of Proposition 7

The first-order conditions for the leader and follower can be written as

$$\begin{aligned}
c_\eta'(\eta_s^k) &= v_{s+1}^k - v_s^k, \\
(c_\lambda^k)'(\lambda_s^k) &= \Phi f'(\lambda_s^k)\, \lambda_{-s}^k\, (v_{s-1}^k - v_s^k) + \Phi\, (v_{s+1}^k - v_s^k), \\
c_\eta'(\eta_{-s}^k) &= v_{-s+1}^k - v_{-s}^k, \\
(c_\lambda^k)'(\lambda_{-s}^k) &= \Phi f(\lambda_s^k)\, (v_{-s+1}^k - v_{-s}^k).
\end{aligned}$$

Using the convex cost forms $c_\eta(\eta) = \frac{1}{\theta}(\alpha\eta)^\theta$ and $c_\lambda^k(\lambda) = \frac{1}{\theta}(\kappa_k\lambda)^\theta$ with $\theta > 1$, we have $c_\eta'(\eta) = \alpha^\theta \eta^{\theta-1}$ and $(c_\lambda^k)'(\lambda) = \kappa_k^\theta \lambda^{\theta-1}$. Hence,

$$\left(\frac{\lambda_{-s}^k}{\eta_{-s}^k}\frac{\eta_s^k}{\lambda_s^k}\right)^{\theta-1} = \frac{(v_{s+1}^k - v_s^k)\, \Phi f(\lambda_s^k)}{\Phi f'(\lambda_s^k)\lambda_{-s}^k(v_{s-1}^k - v_s^k) + \Phi(v_{s+1}^k - v_s^k)} = \frac{f(\lambda_s^k)}{1 + f'(\lambda_s^k)\lambda_{-s}^k\, \dfrac{v_{s-1}^k - v_s^k}{v_{s+1}^k - v_s^k}}.$$

Since $v_{s+1}^k - v_s^k > 0$, $v_{s-1}^k - v_s^k < 0$, $f'(\cdot) > 0$, and $\lambda_{-s}^k \geq 0$, we have

$$f'(\lambda_s^k)\lambda_{-s}^k\, \frac{v_{s-1}^k - v_s^k}{v_{s+1}^k - v_s^k} \leq 0$$

This means that

$$\left(\frac{\lambda_{-s}^k}{\eta_{-s}^k}\frac{\eta_s^k}{\lambda_s^k}\right)^{\theta-1} \geq f(\lambda_s^k) > 1$$

Because $\theta > 1$, this implies

$$\frac{\lambda_{-s}^k}{\eta_{-s}^k}\frac{\eta_s^k}{\lambda_s^k} \;>\; 1 \quad\Longrightarrow\quad \frac{\lambda_{-s}^k}{\eta_{-s}^k} \;>\; \frac{\lambda_s^k}{\eta_s^k}.$$

By the definition of $\psi_\cdot^k = \frac{\lambda_\cdot^k}{\lambda_\cdot^k + \eta_\cdot^k}$, this inequality is equivalent to $\psi_{-s}^k > \psi_s^k$, as claimed.

$\square$

A.2 Proof of Corollary 5

In the case of neck-and-neck competition ($s = 0$), the innovation generality of both firms is

$$\psi_0^k = \frac{\kappa_k^{-\frac{\theta}{\theta-1}} \left[\Phi(v_1^k - v_0^k)\right]^{\frac{1}{\theta-1}}}{\kappa_k^{-\frac{\theta}{\theta-1}} \left[\Phi(v_1^k - v_0^k)\right]^{\frac{1}{\theta-1}} + \alpha^{-\frac{\theta}{\theta-1}} (v_1^k - v_0^k)^{\frac{1}{\theta-1}}}$$
$$= \frac{(\Phi\alpha^\theta)^{\frac{1}{\theta-1}}}{(\Phi\alpha^\theta)^{\frac{1}{\theta-1}} + \kappa_k^{\frac{\theta}{\theta-1}}}.$$

Next, consider the limit as the technology gap $s \to \infty$. From the pair of profits (90), we have $\pi_s \to 1$ and $\pi_{-s} \to 0$. This implies

$$v_s^k \to \frac{1}{r+\tau}, \quad v_{-s}^k \to 0,$$

and hence

$$|v_{-s+1}^k - v_{-s}^k| \to 0, \quad |v_{s-1}^k - v_s^k| \to 0, \quad |v_{s+1}^k - v_s^k| \to 0 \qquad \text{as } s \to \infty.$$

The follower's FOC for general innovations is

$$\kappa_k^\theta (\lambda_{-s}^k)^{\theta-1} = \Phi f(\lambda_s^k)(v_{-s+1}^k - v_{-s}^k).$$

Since the right-hand side vanishes as $s \to \infty$, we have $\lambda_{-s}^k \to 0$. Therefore, the innovation generality of the leader can be written as

$$\psi_s^k = \frac{\kappa_k^{-\frac{\theta}{\theta-1}} \left[\Phi f'(\lambda_s^k)\lambda_{-s}^k(v_{s-1}^k - v_s^k) + \Phi(v_{s+1}^k - v_s^k)\right]^{\frac{1}{\theta-1}}}{\kappa_k^{-\frac{\theta}{\theta-1}} \left[\Phi f'(\lambda_s^k)\lambda_{-s}^k(v_{s-1}^k - v_s^k) + \Phi(v_{s+1}^k - v_s^k)\right]^{\frac{1}{\theta-1}} + \alpha^{-\frac{\theta}{\theta-1}} (v_{s+1}^k - v_s^k)^{\frac{1}{\theta-1}}}.$$

Dividing numerator and denominator by $(v_{s+1}^k - v_s^k)^{\frac{1}{\theta-1}}$ yields

$$\psi_s^k = \frac{\Phi^{\frac{1}{\theta-1}} \kappa_k^{-\frac{\theta}{\theta-1}}}{\Phi^{\frac{1}{\theta-1}} \kappa_k^{-\frac{\theta}{\theta-1}} + \alpha^{-\frac{\theta}{\theta-1}} \left[\frac{f'(\lambda_s^k)\lambda_{-s}^k(v_{s-1}^k - v_s^k)}{v_{s+1}^k - v_s^k} + 1\right]^{-\frac{1}{\theta-1}}}.$$

As $s \to \infty$, we use $|v_{s-1}^k - v_s^k| \to |v_{s+1}^k - v_s^k|$ and the fact that $f'(0)$ is finite, to obtain

$$\psi_s^k \to \frac{(\Phi\alpha^\theta)^{\frac{1}{\theta-1}}}{(\Phi\alpha^\theta)^{\frac{1}{\theta-1}} + \kappa_k^{\frac{\theta}{\theta-1}}}.$$

The leader's FOC for general innovations is

$$\kappa_k^\theta(\lambda_s^k)^{\theta-1} = \Phi f'(\lambda_s^k)\lambda_{-s}^k(v_{s-1}^k - v_s^k) + \Phi(v_{s+1}^k - v_s^k).$$

Since the right-hand side vanishes as $s \to \infty$, it follows that $\lambda_s^k \to 0$. Hence the innovation generality of the follower is

$$\psi_{-s}^k = \frac{\kappa_k^{-\frac{\theta}{\theta-1}}\left[\Phi f(\lambda_s^k)(v_{-s+1}^k - v_{-s}^k)\right]^{\frac{1}{\theta-1}}}{\kappa_k^{-\frac{\theta}{\theta-1}}\left[\Phi f(\lambda_s^k)(v_{-s+1}^k - v_{-s}^k)\right]^{\frac{1}{\theta-1}} + \alpha^{-\frac{\theta}{\theta-1}}(v_{-s+1}^k - v_{-s}^k)^{\frac{1}{\theta-1}}}.$$

Dividing numerator and denominator by $(v_{-s+1}^k - v_{-s}^k)^{\frac{1}{\theta-1}}$ gives

$$\psi_{-s}^k = \frac{\kappa_k^{-\frac{\theta}{\theta-1}}[\Phi f(\lambda_s^k)]^{\frac{1}{\theta-1}}}{\kappa_k^{-\frac{\theta}{\theta-1}}[\Phi f(\lambda_s^k)]^{\frac{1}{\theta-1}} + \alpha^{-\frac{\theta}{\theta-1}}}.$$

Since $f(0) = 1$, letting $s \to \infty$ yields

$$\psi_{-s}^k \to \frac{(\Phi\alpha^\theta)^{\frac{1}{\theta-1}}}{(\Phi\alpha^\theta)^{\frac{1}{\theta-1}} + \kappa_k^{\frac{\theta}{\theta-1}}}.$$

Thus, in the limit as $s \to \infty$, the innovation generality of the leader and follower converge to the same value. □

A.3 Proof of Proposition 8

From Proposition 7, we have $g_s < 0$ for all $s \geq 1$. By construction, we have $g_0 = \psi_0^k - \psi_{-0}^k = 0$.

Suppose, for contradiction, that g_s is not non-monotonic. Since $g_0 = 0 > g_1$, g_s cannot be (strictly) increasing; hence it must be non-increasing on $\mathbb{N}$. Formally, for all $a, b, c \in \mathbb{N}$ with $a < b < c$,

$$g_a \;\geq\; g_b \;\geq\; g_c.$$

In particular, for every $s \geq 1$,

$$g_s \;\leq\; g_1 \;<\; 0,$$

so $\limsup_{s\to\infty} g_s \leq g_1 < 0$. This contradicts Corollary 5, which states $\lim_{s\to\infty} g_s = 0$. Therefore, g_s must be non-monotonic.

Note that market concentration in the model increases strictly with the distance between the leader and the follower.

$$HHI_s = \big(\frac{\rho_s^{1-\sigma}}{\rho_s^{1-\sigma}+1}\big)^2 + \big(\frac{1}{\rho_s^{1-\sigma}+1}\big)^2$$

with ρ_s the price ratio between the leader and the follower, which satisfies $\rho_s^\sigma = \gamma^{-s}\frac{\sigma\rho_s^{\sigma-1}+1}{\sigma+\rho_s^{\sigma-1}}$. Therefore, the innovation generality gap also varies non-monotonically with the level of market concentration. □

A.4 Computation Algorithm

Solving the Balanced Growth Path.

1. Guess the entrants' R&D effort λ_e, the aggregate general R&D efforts Λ and the probability of entering a market of type k, i.e., p_k.
2. The creative destruction rate of market k is given by: $\tau^k = \frac{p_k\Phi\lambda_e}{2\chi_k}$.
3. Solve the incumbents' innovation decisions. Compute and update the aggregate general R&D efforts Λ.
4. Verify the entrants' decisions from (97).
5. Calculate the average growth rate in each market and check if it is equalized from (99).
6. If not converge, update values of λ_e, Λ and $\{p_k\}$, and repeat the process.

General R&D Subsidies.

1. Guess an implied subsidy ν.
2. Repeat the process in "Solving the Balanced Growth Path".
3. Check the budget constraint (113). If not converge, repeat the process.

A.5 Details about Simulated Method of Moments

In Section 4, I apply the simulated method of moments for the internal calibration. In particular, the objective function can be written as:

$$\min_{\upsilon} \sum_{j=1}^{J} w_j \big(\frac{m(\upsilon_j) - \hat{m}(\upsilon_j)}{\hat{m}(\upsilon_j)}\big)^2$$

where υ is the vector of calibrated parameters, $m(\upsilon_j)$ and $\hat{m}(\upsilon_j)$ are the model-generated and empirical moments, respectively. I set the weights w_j such that the moments regarding innovation generality, aggregate growth and entry rates, are weighted 5 times more than other moments.

A.6 Details about the Average Relative Backward Citations

In the model, the average ratio of cross-industry to within-industry backward citations is calculated as:

$$\frac{\sum_k \sum_{s=0}^{\infty} \chi_k (\Phi - 1) \mu_s^k (\lambda_s^k + \lambda_{-s}^k)}{\sum_k \sum_{s=0}^{\infty} \chi_k \mu_s^k \left[f(\lambda_s^k) - 1 \right] \lambda_{-s}^k}$$

which is essentially the ratio of external effect of cross-industry general knowledge spillovers to the within-industry spillovers, weighted by the industry size and firm-level growth.

B Empirical Analysis: Additional Exercises and Robustness

B.1 Detailed Information on Patent 6550058

Figure 42: Detailed Patent Information: Patent ID 6550058

(12) **United States Patent**
Wynn

(10) **Patent No.:** **US 6,550,058 B1**
(45) **Date of Patent:** **Apr. 15, 2003**

(54) **STACK CLEARING DEVICE AND METHOD**

(75) Inventor: **Allen C. Wynn**, Round Rock, TX (US)

(73) Assignee: **International Business Machines Corporation**, Armonk, NY (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/497,606**

(22) Filed: **Feb. 3, 2000**

(51) **Int. Cl.**[7] **G06F 9/45**
(52) **U.S. Cl.** **717/158**; 717/154; 717/130
(58) **Field of Search** 717/152, 157, 717/158, 159, 124, 127, 128, 130, 131, 132, 133, 154; 711/100, 103; 712/202; 713/189, 193, 194, 200, 201, 185

(56) **References Cited**

U.S. PATENT DOCUMENTS

3,889,243	A		6/1975	Drimak	
4,485,435	A		11/1984	Sibley	
4,751,636	A		6/1988	Sibley	
5,193,180	A	*	3/1993	Hastings	717/163
5,293,385	A	*	3/1994	Hary	714/38
5,628,016	A	*	5/1997	Kukol	717/114
6,009,258	A	*	12/1999	Elliott	703/22
6,085,029	A	*	7/2000	Kolawa et al.	714/38
6,189,141	B1	*	2/2001	Benitez et al.	717/153
6,260,187	B1	*	7/2001	Cirne	717/110
6,351,843	B1	*	2/2002	Berkley et al.	709/332
6,385,727	B1	*	5/2002	Cassagnol et al.	713/193
2001/0056539	A1	*	12/2001	Pavlin et al.	713/193

OTHER PUBLICATIONS

"Secure Deletion of Data from Magnetic and Solid–State Memory", Gutmann, Peter, Department of Computer Science, University of Auckland, Jul. 1996, retrieved from http://www.cs.auckland.ac.nz/~pgut001/pubs/secure_del.html, May 21, 2002.*

* cited by examiner

Primary Examiner—Tuan Q. Dam
Assistant Examiner—Mary J. Steelman
(74) *Attorney, Agent, or Firm*—Robert H. Frantz; David A. Mims, Jr.

(57) **ABSTRACT**

A method for removing residual data from a computer program stack prior to returning control to a calling or controlling process with system and method for automatic inclusion thereof into software application programs at the time of production of executable code. Two methods, one for removing residual data from a relatively small stack frame and another for removing residual data from a large stack frame, are automatically inserted into application program code during an enhanced compiling method. Two compiler controls allow a software designer to globally include the stack cleaning feature in all code being produced, or to selectively include the stack cleaning feature into certain indicated modules, code areas, or procedures.

36 Claims, 3 Drawing Sheets

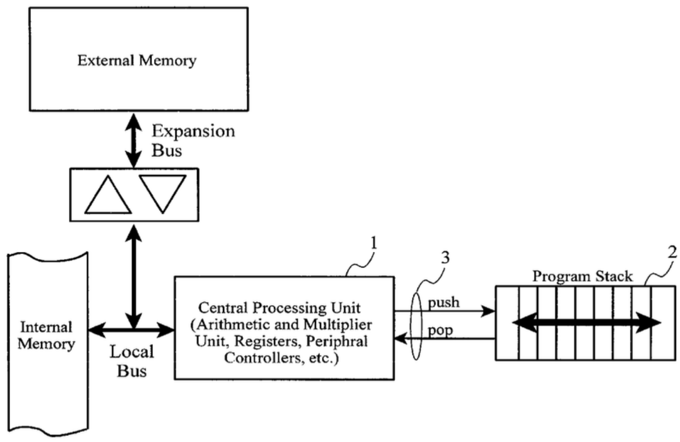


Table 43: Generality Score Construction: Patent 6550058

CPC subclasses	Definitions	Citations
G06F	electric digital data processing	9
G06Q	ICT for admin., comm., fin., manag. or superv. purposes	1

Note: The first and second column list the CPC subclasses the citing patents coming from and their definitions. The third column lists the number of citing patents that belong to the subclass.

B.2 Heterogeneity in Innovation Generality Across Industries.

There exists substantial heterogeneity in innovation generality across industries. Table 44 reports examples of average industry-level patent generality in the United States during 1995–2000.

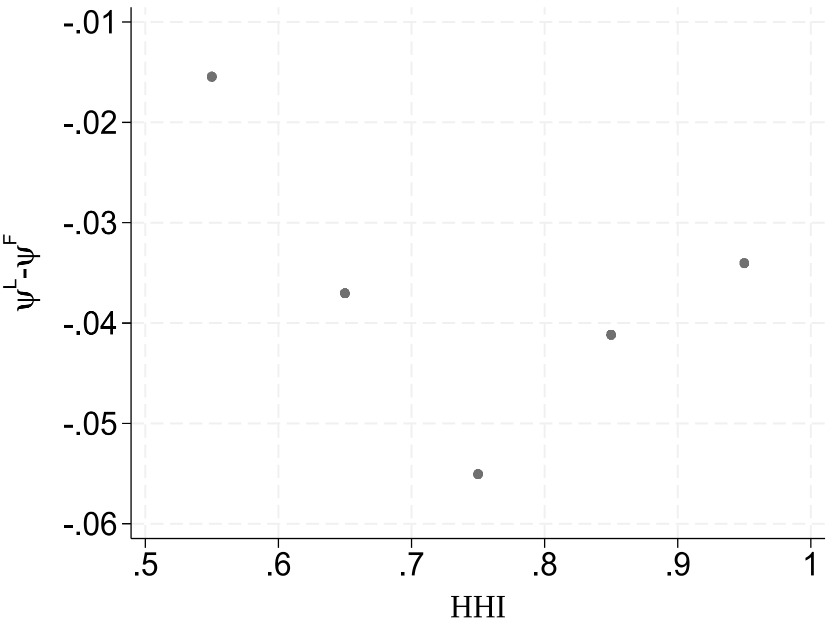
Table 44: Heterogeneity in Innovation Generality Across Industries

SIC code	**Details**	**Avg. Patent Generality**
7323	Credit Reporting Services	0.71
4832	Radio Broadcasting Stations	0.71
1731	Electrical Work	0.70
7331	Direct Mail Advertising Services	0.70
8734	Testing Laboratories	0.67
3443	Fabricated Plate Work	0.31
3221	Glass Containers	0.29
3532	Mining Machinery and Equipment	0.28
5093	Scrap and Waste Materials	0.24
5944	Jewelry Stores	0.06

Note: The first column reports industries' 4-digit SIC codes, the second provides their detailed definitions, and the third presents industry-level patent generality for 1995–2000.

B.3 Empirical Relationship Between Gaps in Innovation Generality and Market Concentration

Figure 43: Empirical Relationship: Gaps in Generality and Market Concentration

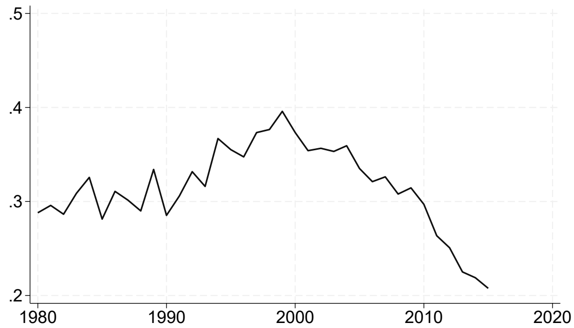


Note: This figure plots the empirical relationship between gaps in innovation generality, i.e., $(\psi^L - \psi^F)$, and market concentration, measured by HHI. The innovation generality gap is taken as the median value in each bin of HHI (with an interval of 0.1). The sample ranges from 1980 to 2015.

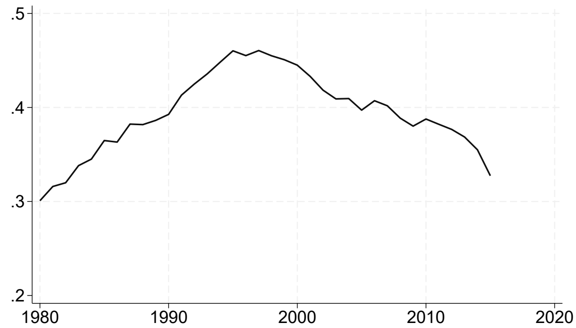
B.4 Average Patent Generality Across Sectors

Figure 44: Average Patent Generality Across Sectors

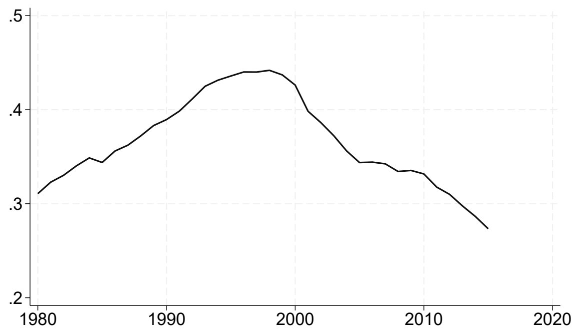
(a) Mining and Construction: SIC Codes 1000-1799



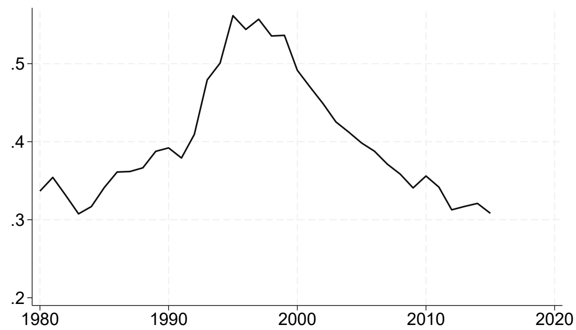
(b) Manufacturing I: SIC Codes 2000-2999



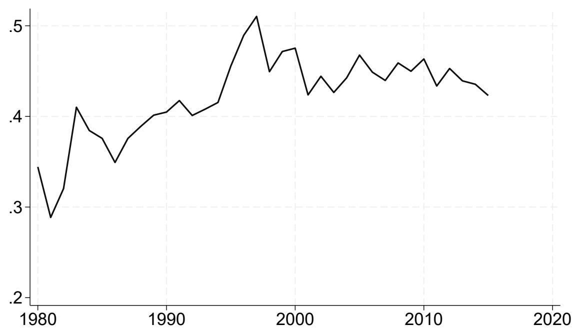
(c) Manufacturing II: SIC Codes 3000-3999



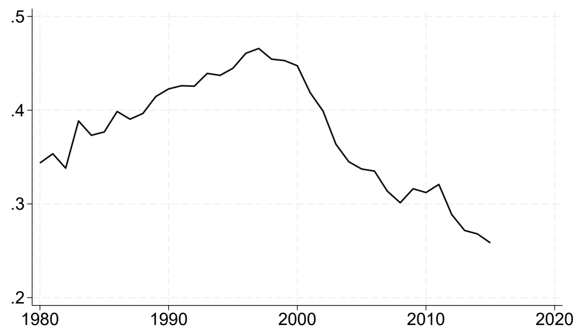
(d) Transportation, Communications, Electric, Gas and Sanitary service: SIC Codes 4000-4999



(e) Trade: SIC Codes 5000-5999



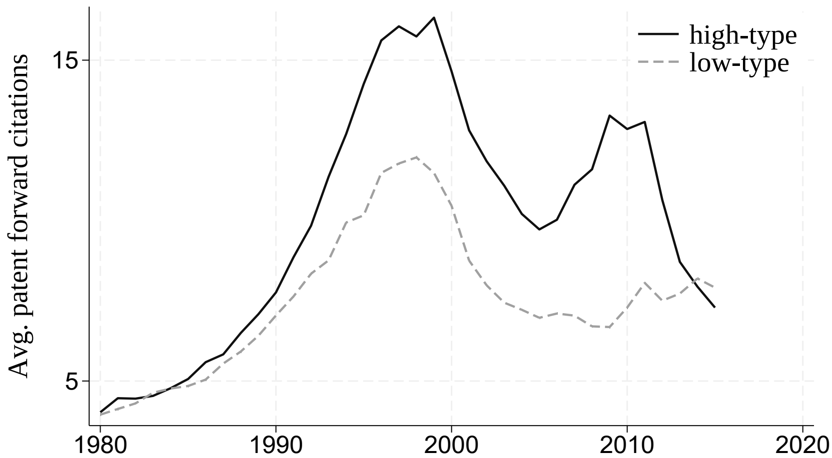
(f) Services: SIC Codes 7000-8999



Note: The figure shows the average patent generality across sector according to one-digit SIC codes, based on citation information within 5-year periods following a patent's grant. SIC codes 2000-2999 generally represent manufacturing industries related to food, textiles, wood, paper, chemicals, and petroleum products, while codes 3000-3999 include fabricated metal products, machinery, electrical equipment, transportation equipment, and instruments.

B.5 Average Patent Forward Citations

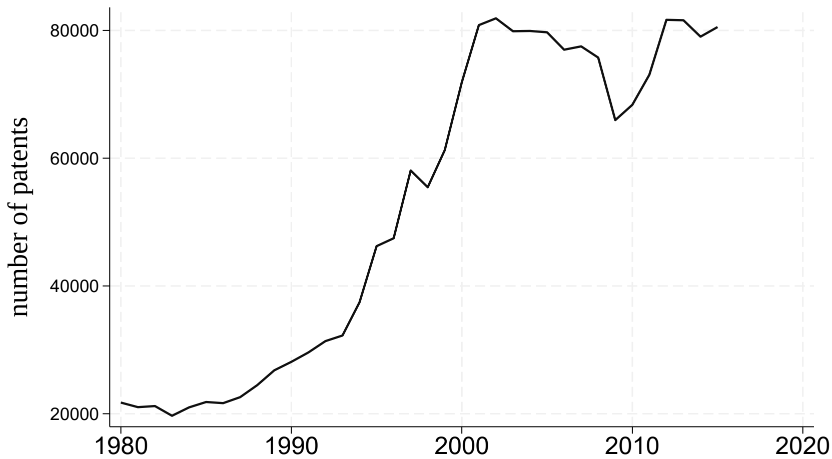
Figure 45: Average Patent Forward Citations (5-year window)



Note: The figure shows the average number of forward citations received within five years of the patent grant for patents in high- and low-type industries from 1980 to 2015. The classification of industry types follows the definitions provided in the main text.

B.6 Total Patent Production: Publicly-listed Firms

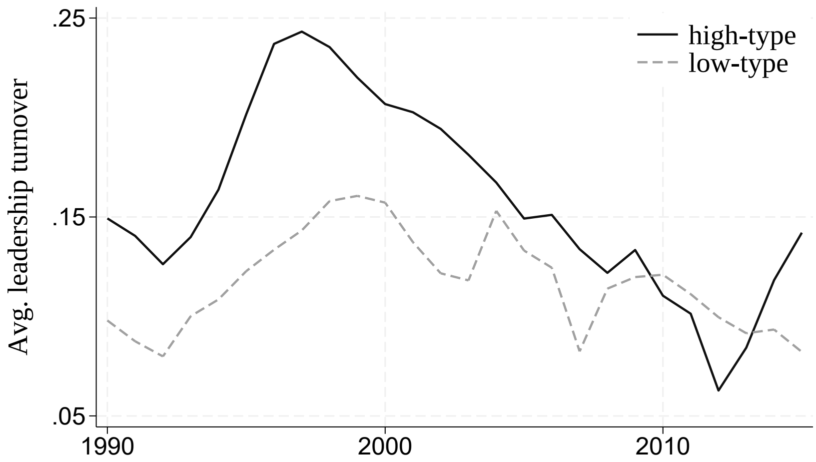
Figure 46: Total Patent Production: Publicly-listed Firms



Note: The figure shows the total number of patents produced by publicly-listed firms from 1980 to 2015. Note that firm sample excludes agricultural (SIC code 0100-0999), utility (SIC code 4900-4949), financial (SIC code 6000-6799), and non-operating establishments (SIC code 9995).

B.7 Average Leadership Turnover

Figure 47: Average Leadership Turnover



Note: The figure shows average leadership turnover rate (3-year moving average) in high- and low-type industries from 1990 to 2015, respectively.

B.8 Robustness Check: Tests of Propositions

Test 1: Definition of Leaders. If leaders are defined as firms whose sales are above the top 5 percentile in their industries, the result of the estimation (106) in the main text becomes:

$$\psi_{f,t} = \underset{(s.e.\ 0.002)}{-\ 0.030} \cdot \mathbb{1}(Leader)_{f,t} + \delta_{j,t} + \epsilon_{f,t}$$

In this specification, holding leadership or not accounts about 0.14 of the standard deviation of firm-level innovation generality.

Test 2: Using Mean Generality Instead of Median. The estimation (107) can be written as:

$$\psi^L_{j,t} - \psi^F_{j,t} = \gamma HHI_{j,t} + \zeta HHI^2_{j,t} + \delta_j + \delta_t + \epsilon_{j,t}$$

Table 45 reports the results under different specifications. In (1), I use the median value of the innovation generality of followers without controlling for the number of firms in the industry. (2) is the baseline estimation in the main text. In (3), I use the mean value of the innovation generality of followers to represent $\psi^F_{j,t}$ in industries with multiple followers, instead of using median value. In (4), I use the mean value and further control for the number of firms in the industry.

Test 2: Measurement of Market Concentration. If instead HHI is constructed using all firms in CRSP-Compustat, regardless of having patent applied or not, the results

Table 45: Robustness Check: Test of Proposition 2

	(1)	(2)	(3)	(4)
	Median	Median	Mean	Mean
γ	-0.221^{**}	-0.357^{***}	-0.210^{**}	-0.314^{***}
	(0.086)	(0.098)	(0.084)	(0.103)
ζ	0.156^{**}	0.229^{***}	0.143^{**}	0.198^{**}
	(0.080)	(0.080)	(0.072)	(0.078)
Number of firms		✓		✓
R-squared	0.135	0.139	0.135	0.139
Observations	5,542	5,542	5,542	5,542

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

are as follows:

$$\psi^L_{j,t} - \psi^F_{j,t} = \underset{(s.e.\ 0.091)}{-\ 0.326} \cdot HHI_{j,t} + \underset{(s.e.\ 0.092)}{0.207} \cdot HHI^2_{j,t} + \delta_j + \delta_t + \epsilon_{j,t}$$

Note that in this case, leaders are defined as firms with the highest sale among all firms that patent within the same industry (4-digit SIC code). The controls are the same as in the main text.

B.9 Robustness Check: Alternative Generality Score Construction

Including patents with zero forward citations. Note that the definition of patent generality in the main text requires a patent to have at least one forward citation within five years of its grant. Among all utility patents assigned to publicly listed U.S. firms between 1980 and 2018, approximately 26.6% have no forward citations within this five-year window. In this analysis, rather than excluding these patents as in the main text, I assign them a generality score of 0.

Test 1: Based on the alternative generality score, the result of the estimation (106) becomes:

$$\psi_{f,t} = \underset{(s.e.\ 0.002)}{-\ 0.034} \cdot \mathbb{1}(Leader)_{f,t} + \delta_{j,t} + \epsilon_{f,t}$$

Test 2: Based on the alternative generality score, the result of the estimation (107) becomes:

$$\psi^L_{j,t} - \psi^F_{j,t} = \underset{(s.e.\ 0.099)}{-\ 0.292} \cdot HHI_{j,t} + \underset{(s.e.\ 0.081)}{0.178} \cdot HHI^2_{j,t} + \delta_j + \delta_t + \epsilon_{j,t}$$

Cost-adjusted firm-level innovation generality. Note that in the main text, firm-level innovation generality is defined as the simple average of patent generality scores. In this analysis, I adjust this measure by incorporating the value of each patent, based on Kogan et al. (2017). The cost-adjusted measure is defined as follows:

$$\widetilde{\psi}_{f,t} = \frac{\sum_{i=1}^{n} \psi_{i,t}/\xi_{i,t}}{\sum_{i=1}^{n} 1/\xi_{i,t}} \tag{201}$$

where $\psi_{i,t}$ denotes the generality score of patent i applied for in year t, and $\xi_{i,t}$ represents the value of patent i, deflated to 1982 (million) dollars using the CPI (Kogan et al., 2017). The basic idea is that if each patent's "cost" (e.g., R&D expenditure or effort) can be approximated by its real value, then this measure reflects on average how general the firm's innovations are, after accounting for how costly they are to produce.

Test 1: Based on the alternative generality score, the result of the estimation (106) becomes:

$$\widetilde{\psi}_{f,t} = \underset{(s.e.\ 0.003)}{-\ 0.026} \cdot \mathbb{1}(Leader)_{f,t} + \delta_{j,t} + \epsilon_{f,t}$$

Test 2: Based on the alternative generality score, the result of the estimation (107) becomes:

$$\widetilde{\psi}^L_{j,t} - \widetilde{\psi}^F_{j,t} = \underset{(s.e.\ 0.101)}{-\ 0.310} \cdot HHI_{j,t} + \underset{(s.e.\ 0.082)}{0.199} \cdot HHI^2_{j,t} + \delta_j + \delta_t + \epsilon_{j,t}$$

B.10 Robustness Check: Definition of Major Changes in Non-compete Enforceability

Table 46 shows the estimation results of equation (108) when a significant change in state-level non-compete enforceability is alternatively defined as a 0.01 shift in the normalized index. Using this definition, I identify 19 major changes between 1991 and 2014, of which 14 involved stronger enforcement.

Table 47 shows the estimation results of equation (108) when a significant change in state-level non-compete enforceability is alternatively defined as a 0.06 shift in the

Table 46: Within-industry Spillovers: Smaller shift

β_N	-0.015^{*}
	(0.008)
$\beta_{N,L}$	0.033^{***}
	(0.009)
Industry-Year FE	✓
State FE	✓
R-squared	0.454
Observations	23,764

Standard errors are clustered at the state level.
*** p<0.01, ** p<0.05, * p<0.1

Note: This table shows the results from estimation (108).

normalized index. Using this definition, I identify 14 major changes between 1991 and 2014, of which 10 involved stronger enforcement.

Table 47: Within-industry Spillovers: Larger shift

β_N	-0.016^{*}
	(0.008)
$\beta_{N,L}$	0.041^{***}
	(0.008)
Industry-Year FE	✓
State FE	✓
R-squared	0.454
Observations	23,764

Standard errors are clustered at the state level.
*** p<0.01, ** p<0.05, * p<0.1

Note: This table shows the results from estimation (108).

B.11 Robustness Check: Pre-determined Leadership

Table 48 shows the estimation results of equation (108) when the leadership indicator in the interaction term is constructed using pre-determined leader status measured one or two periods prior to the treatment year. This is to avoid treatment bias: If policy affects competition dynamics, leadership could shift as a result of the reform.

Table 48: Within-industry Spillovers: Pre-determined Leadership

	$\mathbb{1}(Leader)_{f,\tau-1}$	$\mathbb{1}(Leader)_{f,\tau-2}$
β_N	-0.015^{*}	-0.015^{*}
	(0.008)	(0.008)
$\beta_{N,L}$	0.030^{***}	0.032^{***}
	(0.010)	(0.012)
Industry-Year FE	✓	✓
State FE	✓	✓
R-squared	0.454	0.454
Observations	23,764	23,764

Standard errors are clustered at the state level.
*** p<0.01, ** p<0.05, * p<0.1

Note: This table shows the results from estimation (108).

B.12 Robustness Check: Persistent Leadership

Table 49 shows the estimation results of equation (108) when the leadership indicator in the interaction term is based on a measure of persistent leadership, defined as maintaining leader status from the time of the reform (i.e., τ) through period t. This approach accounts for potential leadership turnover over time.

Table 49: Within-industry Spillovers: Persistent Leadership

β_N	-0.016^{**}
	(0.008)
$\beta_{N,L}$	0.044^{***}
	(0.008)
Industry-Year FE	✓
State FE	✓
R-squared	0.454
Observations	23,764

Standard errors are clustered at the state level.
*** p<0.01, ** p<0.05, * p<0.1

Note: This table shows the results from estimation (108).

B.13 Additional Stylized Facts

Fact 3: Entrants have higher innovation generality.

I show entrants have higher innovation generality compared with incumbents using the following estimation:

$$\psi_{f,t} = \underset{(s.e.\ 0.002)}{0.033} \cdot \mathbb{1}(Entrant)_{f,t} + \delta_{j,t} + \epsilon_{f,t} \tag{202}$$

where $\psi_{f,t}$ represents the average patents generality of firm f in year t. $\mathbb{1}(Entrant)_{f,t}$ is an indicator function, which equals one if firm f is an entrant (i.e., within 5 years since first appeared in Compustat) in industry j in year t. $\delta_{j,t}$ captures industry-year fixed effects.

Fact 4: General patents have lower self-cite rates.

Furthermore, I study how general knowledge produced by a firm relies on firm's own knowledge pool from the following estimation:

$$s_{i,t} = \underset{(s.e.\ 0.001)}{-\ 0.031} \cdot \psi_i + \delta_f + \delta_t + \epsilon_{i,t} \tag{203}$$

where $s_{i,t} \in [0,1]$ is patent i's self-cite rate. Self-cite rate is the share of backward citations made to patents with the same assignee (i.e., same firms) (Akcigit and Kerr, 2018). Patents having higher self-cite rates are more dependent on firm's own technology pools, and hence are more firm-specific in nature. As before, ψ_i measures generality of patent i. δ_f and δ_t are firm fixed effects and application year fixed effects. Patent-level generality explains about 0.14 of the standard deviation in the self-cite rate. This suggests that general knowledge relies less on firm-specific knowledge.

B.14 Robustness Check: Generality and Market Concentration

Measurement of Market Concentration. Following estimation (111) and (112) in the model validation section, I use HHI constructed using all firms in CRSP-Compustat, regardless of having patent applied or not, the results are as follows:

$$\psi^L_{j,t} = \underset{(s.e.\ 0.060)}{-\ 0.124} \cdot HHI_{j,t} + \underset{(s.e.\ 0.057)}{0.088} \cdot HHI^2_{j,t} + \delta_j + \delta_t + \epsilon_{j,t} \tag{204}$$

$$\psi^F_{j,t} = \underset{(s.e.\ 0.069)}{0.229} \cdot HHI_{j,t} - \underset{(s.e.\ 0.069)}{0.191} \cdot HHI^2_{j,t} + \delta_j + \delta_t + \epsilon_{j,t} \tag{205}$$

Controlling for Citation Counts. If I further control for the log number of average forward citations (within 5-year window since the patent grant) received by patents produced by firm f in year t, the results are as follows:

$$\psi^L_{j,t} = -\underset{(s.e.\ 0.058)}{0.153} \cdot HHI_{j,t} + \underset{(s.e.\ 0.047)}{0.116} \cdot HHI^2_{j,t} + \delta_j + \delta_t + \epsilon_{j,t} \tag{206}$$

$$\psi^F_{j,t} = \underset{(s.e.\ 0.071)}{0.164} \cdot HHI_{j,t} - \underset{(s.e.\ 0.058)}{0.092} \cdot HHI^2_{j,t} + \delta_j + \delta_t + \epsilon_{j,t} \tag{207}$$