Self-Selection and the Diminishing Returns of Research*

Lorenz K.F. Ekerdt
Center for Economic Studies, U.S. Census Bureau

Kai-Jie Wu
Penn State

November 2024

[click for latest version]

Abstract

The downward historical trend of research productivity has been used to suggest that there are severe permanent diminishing returns of knowledge production. We argue that a substantial portion of the trend is a transitional composition effect resulting from self-selection in researchers' ability and the expansion of the researcher sector. We quantify said effect with a semi-endogenous growth model in which workers self-select into research together with microdata on sectoral earnings distributions. Our results suggest that the average ability of researchers has fallen substantially. We then revisit the estimation of the knowledge production function and its resulting prediction on long-run economic growth. We find that separating transitional diminishing returns from permanent ones nearly doubles the long-run growth rate of per capita income predicted by a broad class of growth models.

JEL codes: D24, E23, J24, O47.

*We thank our adviser Mark Bils for his invaluable guidance and good humor throughout this project. We have also benefited greatly from discussions with George Alessandria, Yan Bai, Travis Baseler, Richard Beem, Corina Boar, Melissa Chow, Charles Jones, Lisa Kahn, Zachary Kroff, Narayana Kocherlakota, Ronni Pavan, John Singleton, Martha Stinson, Sean Wang, and seminar participants at the University of Rochester and the Center for Economic Studies. Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Census Bureau.

1 Introduction

The creation of ideas is the engine of economic growth. The modern era has seen an enormous investment of labor and resources into the forward motion of this machine: expenditure on intellectual property products in the U.S. has increased 23-fold since the 1930s (Bloom, Jones, Van Reenen, and Webb, 2020) and the expansion of the workforce devoted to research and development (R&D) has been similarly dramatic.[1] Given this rapid expansion, we would expect to see that growth has accelerated; yet by most measures, the aggregate growth rate of total factor productivity (TFP) in the U.S. has been flat or declining. The unavoidable conclusion of the past 8 decades of growth is thus that the productivity of research has fallen dramatically. Is the cause of this decline fundamental to the accumulation of ideas? Will this build-up of friction continually slow down the forward motion of growth, eventually resulting in a world with low or zero growth rates?

In this paper, we argue that a substantial portion of the declining productivity of research is the product of a transitional composition effect instead of a permanent phenomenon. We show that the share of the labor force employed in research has increased 3-fold since the 1960s. If workers are sorted into sectors according to ability, this rapid expansion cannot have been achieved without changing the average ability among researchers. If research is positively selected on ability, i.e., workers with a comparative advantage in research also tend to have an absolute advantage in research, the direction of this change is unambiguously negative. This source of diminishing returns is transitional because it is expected to fade away when the share of workers employed in research ceases to rise.

To explore this line of reasoning, we marshal microdata on sectoral choices and earnings to estimate a semi-endogenous growth model with Roy-like selection into research. Our estimated model suggests that average ability in the U.S. research sector has fallen by about 49% since 1960. We gauge the importance of this decline by discussing its implication on estimating the permanent diminishing returns of research, as captured by the scale elasticity of existing knowledge in creating new knowledge. Permanent diminishing returns play a key role in predicting long-run economic growth in modern growth models. Previous studies tend to understate the long-run growth rate because they attribute all observed declines in research productivity to permanent diminishing returns in research (e.g., Jones, 1995, 2002; Bloom et al., 2020). We find that separat-

[1] Bloom et al. (2020) deflates expenditure on intellectual property products by the nominal wage of males with a bachelor's or higher degree as a measurement of the "effective number of researchers."

ing the proposed source of transitional diminishing returns from the observed decline of research productivity implies 1.85 times higher long-run growth rates in a broad class of growth models. When seen through the lens of a simple semi-endogenous growth model, the long-run growth rate of income increases by 0.5 percentage points. We conclude that not only ideas are getting harder to find; researchers are too.

Our focus on the earnings distribution of researchers leverages unique information about the work activities of college-educated workers provided by the U.S. National Survey of College Graduates (NSCG). We define researchers as workers whose occupations are categorized as scientists or engineers and who report their work activity as R&D. The NSCG enables us to observe researchers' earnings for each survey year between 1992 to 2017 in the U.S. To obtain a longer trend of the research labor force and their earning distribution, we impute the occupational-level information in the NSCG to the Decennial Census and the American Community Survey (ACS) to construct aggregate moments over the period 1960-2017. Our empirical findings confirm the fact that the overall share of researchers in the U.S. labor force has been growing over time. Perhaps more strikingly, we find that the share of researchers among the college-educated labor force has been declining. This suggests that the increase in researchers, in conjunction with a fall in their average productivity, is largely associated with the dramatic rise in college attendance.

Motivated by these trends, our model of researcher supply features both educational and sectoral choices. Individuals are heterogeneous in three dimensions: (i) their expected return on a college education, (ii) their ability in the research sector, and (iii) their ability in the non-research college-graduate sector. Given their type, individuals decide whether to go to college and whether to pursue research. The model naturally implies that forces expanding the research sector will generate changes in the average ability of researchers. But the direction and strength of this impact hinge crucially on whether comparative advantage and absolute advantage align among researchers. In other words, are these added researchers as productive as those who would have been selected into the research sector absent that sectoral expansion?

To answer this question we make use of moments of the sectoral earnings distribution, which speak strongly to the latent ability distribution and hence the direction and magnitude of self-selection in abilities (Heckman and Sedlacek, 1985). We emphasize two features of the earnings data. First, the earnings dispersion of researchers and college workers has increased markedly since 1960. This finding is consistent with positive ability sorting into college education—those marginally selected into college are less productive than those with strong comparative advantage, regardless of pursuing research or not.

Second, longitudinal moments constructed from the panel component of the NSCG reinforce this interpretation. The logic is as follows: assuming that the marginal worker in a sector is more likely to move between sectors, comparing the mean log earnings of workers transiting between sectors to the mean log earnings of those who do not allows us to compare the marginal worker's ability to the average worker's ability. Implementing this comparison in the NSCG, we find that workers transiting from the research sector to the non-research sectors have lower earnings than researchers who do not; through the lens of our model, this can only happen if the marginal researcher has lower ability than the average researcher. Conversely, workers transiting from the non-research sector to the research sector have higher earnings than non-researchers who do not, suggesting that the marginal non-researcher has higher ability than the average non-researcher.[2]

We incorporate these empirical findings and estimate our model by indirect inference. Specifically, we target our estimation to the following three sets of moments: (i) sectoral employment shares in 1960, (ii) changes in sectoral employment shares and the sectoral variance of log earnings between 1960 and 2017, and (iii) the mean log earnings of workers transiting between the research and non-research sectors relative to the mean log earnings of those who do not.

To make progress, we parameterize the ability distribution with a log-normal form. A vulnerability of this assumption is that, when only estimated using cross-sectional earnings, results are sensitive to the choice of parameterization.[3] Our use of longitudinal moments in the estimation is key towards avoiding this critique because panel data allow for local nonparametric identification of the distribution of ability (see Theorem 11 of Heckman and Honoré, 1990), where local here refers to the domain of relative sectoral returns for which panel data is available. We thus view our longitudinal moments as providing a reasonably robust characterization of the distribution of ability for the years in which we have panel data, whereas cross-sectional moments play the secondary role of extrapolating this distribution to years for which we do not.

Our estimation results suggest that the average ability of researchers in the U.S. has decreased by 49% over the period 1960-2017 due to the transitional composition effect. The main source of the composition effect is the substantial expansion of college education; the magnitude of this channel is mainly inferred from the comovement of earnings dispersion and sectoral

[2] Note that the longitudinal moments only show negative self-selection into the non-research sector conditional on having a college degree. The self-selection into college education can still be positive.

[3] It is possible to estimate a log-normal Roy model using higher order moments of the cross-sectional distribution of log earnings, e.g. skewness as in French and Taber (2011). Other than the downside that this is only valid under log normality, there is no economic intuition for the role played by these higher-order moments in the estimation; more practically, for our purposes, it is unlikely that these higher-order moments are well-measured in our data set.

shares in our estimation. In contrast, self-selection between researchers and non-researchers among college graduates has slightly driven up the average ability of researchers. This is because the share of researchers among college graduates has declined and because abilities are positively self-selected into the research sector, which is mainly inferred from the longitudinal moments in our estimation.

We then re-visit the idea production function and examine the resulting prediction on long-run economic growth. We find that the predicted growth rate in a broad class of semi-endogenous growth models increases by 1.85 times when taking into account the decrease in average ability due to self-selection. For instance, the predicted growth rate in Jones (1995)'s model increases by 0.5 percentage points. This means that, on the balanced growth path, the ratio in per capita incomes between these two hypothetical economies would approximately double every 138 years.

The paper proceeds as follows. In Section 2 we describe the supply-side of a semi-endogenous growth model featuring Roy-like selection into research.[4] We clarify the distinction between permanent diminishing returns to research and the transitional diminishing returns that we estimate in this paper. We also show how our evidence can be used to refine predictions about long-run growth rates. In Section 3, we describe our data sources and measurement of researchers' labor force and earnings, and we discuss aggregate time trends and longitudinal moments of the U.S. researcher labor market. Though our primary focus is the United States, we also show that the trends documented in this paper generally hold in developed countries. In Section 4 we estimate our model using the data moments from Section 3. Section 5 then uses the composition effect inferred by the estimated model to revisit the long-run predictions in growth models. Section 6 concludes.

Related literature. Our closest antecedent in the literature is Bloom et al. (2020), who compile an impressive body of evidence to argue that research productivity is falling across many different areas and levels of detail of the economy. We are also closely related to Jones (2002) and Fernald and Jones (2014), who argue that the recent history of growth in developed countries is best described as a transitional path resulting in a (nearly) constant growth rate rather than a balanced growth path in which various aggregates grow in proportion.

Knowing the strength of self-selection is crucial to assessing the economic impact of such policies as R&D and education. Akcigit, Pearce, and Prato (2020) argue that, by incorporating

[4] As will be shown, the demand-side is not strictly necessary for either our estimation or its implication for long-run growth; for completeness's sake, we describe the demand side of the model and its balanced growth path in Appendix A.

positive self-selection into research, R&D subsidies are less effective than in standard models and that their effect can be strengthened by education subsidies. Akcigit, Pearce, and Prato also provides evidence suggesting positive self-selection into research. Using Danish data, they show that the expansion of Ph.D. enrollment (due to a policy intervention in 2002) is associated with a substantial drop in the average IQ of Ph.D. enrollees.

Methodologically, we draw upon a vast literature in labor economics studying the implications and empirical fit of models of self-selection, particularly Heckman and Sedlacek (1985, 1990), and Heckman and Honoré (1990). We are also closely aligned with recent literature studying the implications of self-selection in the labor market on sectoral and aggregate productivity: Lagakos and Waugh (2013) use a two-sector model to study cross-country productivity differences in agriculture; Young (2014) uses an empirical approach to argue that productivity in the service sector is mismeasured; Hsieh, Hurst, Jones, and Klenow (2019) focus on the productivity-enhancing effect of falling discrimination.

2 Model

We embed a simple selection framework into an otherwise standard semi-endogenous growth model, drawing upon the work of Romer (1990) and Jones (1995). The model features educational choice in the spirit of Willis and Rosen (1979) and sectoral choice in the spirit of Roy (1951). There are two sectors: an idea producing R&D sector and a final goods production sector. The only factor of production is labor, which is segmented into high-skill and low-skill labor. R&D production only uses high-skill labor, whereas production uses both high- and low-skill labor. In order to be employed as a high-skill worker, workers must obtain a college degree. Workers thus decide whether or not to obtain a college degree and, conditional on getting a college degree, whether or not to work as a researcher.

Workers are heterogeneous in their sector-specific high-skill endowments of efficiency units, which we will often refer to as their sector-specific ability. Note that this is the only source of heterogeneity across workers. We take the joint distribution of these endowments as a primitive of the model, and are thus holding it fixed in our estimation. Workers know their sector-neutral ability at the point of choosing whether or not to obtain a college degree, but do not necessarily know their sector-specific abilities. Time is continuous, and endowments are redrawn at each point in time. We are therefore assuming time-specific cohorts, and consequently will match the model to long changes (1960-2017) in the data so that the cohorts do not overlap much.

In the main text, we take efficiency wage rates as given; as will become clear, we focus on a supply-side characteristic of the R&D sector that can be discussed without reference to the demand-side. Nevertheless, we detail the demand-side, and discuss the balanced growth path of the full model, in Appendix A.

With some abuse of terminology, we will often refer to the three different types of labor as sectors, though low-skilled and non-researcher high-skilled labor are both employed in the production of the final good. In what follows, we will use an R subscript to denote model elements relating to the researcher sector, H to denote model elements relating to high-skill production labor, and N to denote model elements relating to low-skill production labor.

2.1 Sectoral production functions

R&D produces ideas according to the idea production function

$$\dot{A}_t = A_t^{\phi} Z_{R,t}, \tag{1}$$

with $\phi < 1$, $\dot{A}_t$ the flow of ideas, A_t the stock of ideas, and $Z_{R,t}$ the total efficiency units of research labor supplied. The elasticity ϕ captures the returns to scale of existing knowledge. The productivity of research depends on the accumulated stock of knowledge due to knowledge spillovers. When $\phi > 0$, higher levels of existing knowledge elevate the productivity of research. This captures the logic of Newton's famous aphorism: contemporary researchers are "standing on the shoulders of giants." The case of $\phi < 0$, on the other hand, can be thought of as "fishing out"—the more ideas that have been discovered, the more difficult it is to find new ones.

The final goods sector admits a representative firm whose production function is:

$$Y_t = A_t G(Z_{H,t}, Z_{N,t}), \tag{2}$$

with Y_t denoting total output, $Z_{H,t}$ the total efficiency units of high-skilled production labor, $Z_{N,t}$ total low-skilled production labor, and G a constant-returns-to-scale aggregator.[5]

[5] When solving for the balanced growth path in Appendix A, we will further restrict the form of this aggregator; however, all results in the main text hold for any aggregator satisfying constant returns to scale.

2.2 Labor supply

A continuum of size L_t workers are endowed with sector-neutral high-skill efficiency units $e^{z_{Ci}}$ and sector-specific high-skill efficiency units $(e^{z_{Ri}}, e^{z_{Hi}})$, with (z_{Ci}, z_{Ri}, z_{Hi}) drawn according to some time-invariant distribution function F. Unless otherwise noted, all expectations in this section are taken across individuals with respect to this distribution. Worker's earnings if employed as a high-skill worker in sector $j \in \{H, R\}$ are $e^{w_{j,t}+z_{Ci}+z_{ji}}$ with $w_{j,t}$ the log sectoral efficiency wage rate. Their utility if employed in high-skill sector j is given by:

$$U_{j,t} = e^{\mu_{j,t}+w_{j,t}+z_{Ci}+z_{ji}} \tag{3}$$

where $\mu_{j,t}$ captures the non-pecuniary returns per efficiency unit in sector $j \in \{H, R\}$. In order to be employed as a high-skilled worker, individuals must obtain a college degree at per efficiency unit utility cost $e^{\mu_{C,t}}$. If they choose not to obtain a college degree, they are employed as a low-skilled worker in the production sector. We assume that all workers have the same endowment of low-skilled efficiency units, which we normalize to 1. A worker's utility if employed as a low-skilled worker is thus

$$U_{N,t} = e^{w_{N,t}+\mu_{N,t}}. \tag{4}$$

Workers make labor supply decisions to maximize utility. Given the form of the utilities, the optimal labor supply decisions are characterized by simple cutoff rules. Let $\tilde{w}_{j,t} = w_{j,t} + \mu_{j,t} - w_{N,t} - \mu_{N,t}$ denote log of the per-efficiency-unit return of working as a high-skilled worker in sector $j \in \{H, R\}$ relative to working as a low-skilled production worker. Then workers will choose to get a college degree if

$$0 \leq z_{Ci} + \mathbb{E}\left(\max\{\tilde{w}_{R,t} + z_{Ri}, \tilde{w}_{H,t} + z_{Hi}\} \mid z_{Ci}\right) - \mu_{C,t}, \tag{5}$$

where we emphasize that workers know their z_{Ci} when choosing whether to obtain a college degree but may or may not know their (z_{Ri}, z_{Hi}).[6] Conditional on obtaining a college degree, they choose to become researchers if

$$\tilde{w}_{H,t} + z_{Hi} \leq \tilde{w}_{R,t} + z_{Ri} \tag{6}$$

[6] All results in this section hold whether workers know their (z_{Ri}, z_{Hi}) when choosing whether to attend college or if they instead form rational expectations of their (z_{Ri}, z_{Hi}) given knowledge of z_{Ci} and the joint distribution F.

Define $z^m_{C,t}$ to be the value of z_{Ci} such that equation (5) holds with equality, and similarly define $z^m_{R,t}(z_H)$ to be the value of z_{Ri} such that equation (6) holds with equality. Intuitively, $z^m_{C,t}$ is the sector-neutral ability of the worker that is on the margin between attending college and not. This value changes over time because it depends on the relative sectoral returns.[7] Similarly, $z^m_{R,t}(z_{Hi})$ is the research-specific ability of the college-educated worker that is on the margin between becoming a researcher and not; this value again changes over time with the relative sectoral returns, and it varies across different values of the high-skill-production-specific ability z_{Hi}. Given these definitions, we can write the total number of efficiency units of research labor supplied to the R&D sector given relative sectoral returns as:

$$Z_{R,t} = L_{R,t} \times \mathbb{E}\left[e^{z_{Ci}+z_{Ri}} \,\middle|\, z_{Ci} \geq z^m_{C,t}, z_{Ri} \geq z^m_{R,t}(z_{Hi})\right]. \tag{7}$$

Changes in the relative returns of working in different sectors change the values of $z^m_{C,t}$, $z^m_{R,t}(z_H)$ and thereby the composition of workers in those sectors; we refer to the resulting changes in the conditional expectation in equation (7) as changes in ability stemming from self-selection.

2.3 Key properties

2.3.1 The scale elasticity of knowledge determines long-run growth

As summarized in Jones (2005), the scale elasticity of knowledge ϕ fundamentally determines the long-term growth rate on a balanced growth path. To see this, express equation (1) as

$$\ln\left(\frac{\dot{A}_t}{A_t}\right) = (\phi - 1)\ln A_t + \ln Z_{R,t} \tag{8}$$

and take derivatives of both sides of this equation with respect to time; noting that the derivative of the left-hand side must be 0 on a balanced growth path and rearranging, we obtain

$$g_A^{\text{BGP}} = \left(\frac{1}{1-\phi}\right) g_{Z_R}^{\text{BGP}} = \left(\frac{1}{1-\phi}\right) g_L, \tag{9}$$

with g_L denoting the growth rate of the labor force. Note that, on the balanced growth path, the share of researchers in the labor force reaches a constant, so there is no change in the average researcher ability due to the composition effect, and the growth rate of researcher efficiency-units

[7] Note that $z^m_{C,t}$ may depend on the worker's (z_{Ri}, z_{Hi}) if they have full knowledge of them when choosing whether to attend college.

equals population growth, $g_{Z_R}^{\text{BGP}} = g_L$.

The intuition behind equation (9) is straightforward: on any balanced growth path with constant population growth, diminishing returns to research (i.e. $\phi < 1$) imply that constant idea growth can only be achieved by constant growth in R&D inputs.[8] Absent other exogenous sources of factor growth, per capita output growth is consequently proportional to population growth. The lower the value of ϕ, the stronger the diminishing returns, the less idea growth given a rate of population growth and thus the lower the per capita output growth rate. We therefore refer to the scale elasticity as the "permanent" diminishing returns to research. Note that equation (9) holds no matter how $g_{z_R}^{\text{BGP}}$ is determined and is thus a statement about any growth model whose aggregate idea production function can be expressed as equation (1).

Within the class of semi-endogenous growth models, different values of ϕ can lead to a wide range of long-run income growth rates. For instance, in (Jones, 1995), the growth rate of income per person is directly given by the ratio: $g_y = g_L/(1-\phi)$. In general, semi-endogenous growth models predict g_y to be proportional to $g_L/(1-\phi)$ in the long run. Assuming a 1% long-run annual population growth rate, the predicted income growth rate for $\phi = -2$ is 0.33%, whereas it increases to 2% when $\phi = 0.5$. This difference can create wide divergences in income levels—the ratio of income levels predicted by $\phi = 0.5$ relative to that predicted by $\phi = -2$ doubles every 42 years.[9]

The scale elasticity is also a key determinant of the speed of convergence to the balanced growth path; fixing research intensity, the half life of convergence to the balanced growth rate given some initial rate is again proportional to $1/1-\phi$: see Section 2.5 in Jones (2022). Similar statements are true if research intensity is not fixed on the transition path: see Atkeson and Burstein (2019).

2.3.2 Self-selection leads to changes in average researcher ability

The key property of the labor supply framework is the evolution of average efficiency units of research supplied per researcher over time. Note first that as long as the relative returns $\tilde{w}_{R,t}$, $\tilde{w}_{H,t}$ do not change, the average ability of researchers does not change. This is simply because the distribution of high-skill efficiency units $(e^{z_{Ci}}, e^{z_{Ri}}, e^{z_{Hi}})$ does not change over time; formally, it

[8] The case of $\phi = 1$, i.e. no diminishing returns, is the one originally considered by Romer (1990).

[9] Jones (2002) considers three values -2.03, -4, and -19; Fernald and Jones (2014) uses -1.63; and Bloom et al. (2020)'s accounting from aggregate TFP suggests -2.1. Peters (2019) on the other hand, estimates ϕ to be 0.88 using a spatial growth model together with data from Germany's post-war population expulsions.

can be seen by noting that the term after L_t in equation (7) is only a function of time through the relative returns $\tilde{w}_{R,t}$ and $\tilde{w}_{H,t}$.

If, on the other hand, the size of the research sector as a share of total labor changes over time, then average researcher ability will generally change, with the rate of change determined by the shape of the joint distribution of worker types. More precisely, the average ability of researchers is decreasing over time whenever the marginal researcher drawn into the sector is worse than the average researcher i.e. if

$$\frac{\dot{Z}_{R,t}}{\dot{L}_{R,t}} < \frac{Z_{R,t}}{L_{R,t}}. \tag{10}$$

In our model, the relative size of the research sector can expand in two ways: by drawing college-educated workers from the production sector to the research sector ($\dot{\tilde{w}}_{R,t} > \dot{\tilde{w}}_{H,t}$), and by inducing low-skilled workers to get a college degree and become researchers ($\dot{\tilde{w}}_{R,t} + \dot{\tilde{w}}_{H,t}) > 0$. There are thus two relevant marginal workers: the worker on the margin between high-skilled production and research, and the worker on the margin between low-skilled production and research.

To illustrate how workers on the margin are key to pinning down the direction and magnitude of self-selection, we look at the average log ability of researchers.[10] Specifically, let

$$\bar{z}_t^R = \mathbb{E}\left[z_{Ci} + z_{Ri} \mid z_{C,t}^m \le z_{Ci}, z_{R,t}^m(z_{Hi}) \le z_{Ri}\right] \tag{11}$$

be average log ability in the research sector. Now suppose for the sake of illustration that the sector-neutral log ability z_{Ci} is independent of the sector-specific log abilities (z_{Ri}, z_{Hi}).[11] Then, taking a time derivative, we can write:

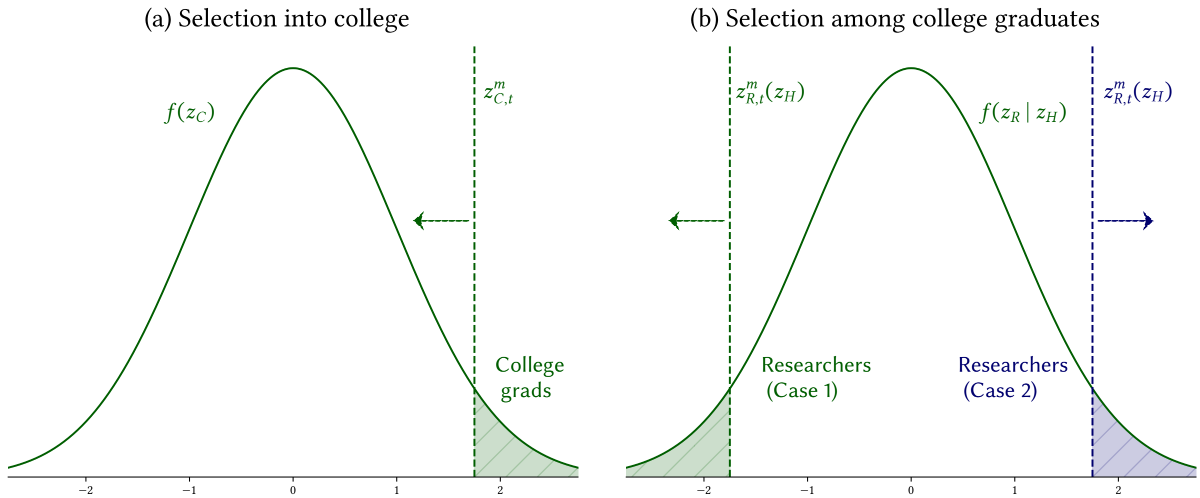
$$\begin{aligned} \dot{\bar{z}}_t^R \; = \; & \underbrace{\left(\dot{\tilde{w}}_{R,t} + \dot{\tilde{w}}_{H,t}\right)\left[\mathbb{E}\left(z_{C,t}^m\right) - \mathbb{E}\left(z_{Ci} \mid z_{Ci} > z_{C,t}^m\right)\right]}_{\text{selection into college}} \\ & + \underbrace{\left(\dot{\tilde{w}}_{R,t} - \dot{\tilde{w}}_{H,t}\right)\left(\mathbb{E}\left[z_{R,t}^m(z_{Hi})\right] - \mathbb{E}\left[z_{Ri} \mid z_{Ri} > z_{R,t}^m(z_{Hi})\right]\right)}_{\text{selection among college graduates}}. \end{aligned} \tag{12}$$

This expression just decomposes the evolution of average log ability among researchers into two

[10] Recall that the precise definition of self-selection is defined by equation (7): $\mathbb{E}\left[e^{z_{Ci}+z_{Ri}} \mid z_{Ci} \ge z_{C,t}^m, z_{Ri} \ge z_{R,t}^m(z_{Hi})\right]$. Here we simply look at the average log ability for illustrative purposes. The cross-time change of average ability among researchers has the same sign as that of log average ability among researchers, and thus the discussion below applies equally to average ability.

[11] Independence between z_{Ci} and (z_{Ri}, z_{Hi}) implies that these two margins do not interact; the corresponding expression without independence can be found in Appendix A.2.

Figure 1: Examples of self-selection into research



Notes: Panel (a): Example of a case when selection into college decreases the average ability among researchers. The green curve depicts the density of the sector-neutral ability z_{Ci}, and the dashed line denotes the lowest z_{Ci} for which a worker chooses to attend college. Panel (b): Example of two cases when selection into research among college graduates changes the average ability among researchers. The green curve depicts the density of the sector-neutral ability $z_{Ri} \mid z_{Hi}$, and the dashed line denotes the lowest z_{Ri} for which a college graduate with high-skill production ability z_{Hi} chooses to become a researcher. Case 1 depicts negative self-selection into research; Case 2 depicts positive self-selection in to research.

components: changes in sector-specific ability z_{Ri} and changes in sector-neutral ability z_{Ci}. We refer to the former as selection among college graduates and the latter as selection into college.

Consider first selection into college: if researchers are expanding as a share of the labor force, i.e. $\dot{\hat{w}}_{R,t}+\dot{\hat{w}}_{H,t} > 0$, selection into college decreases the average log ability among researchers if the marginal college graduate has lower sector-neutral ability than the average researcher, and vice versa. Assuming that workers have no information about their sector-specific abilities (z_{Ri}, z_{Hi}), Panel (a) of Figure 1 illustrates this case. As the return of becoming a college graduate increases, the sector-neutral ability of the marginal college graduate declines, and thus the average sector-neutral ability among researcher also declines.

Selection among college graduates is similar: in the empirically-relevant case in which researchers as a share of college graduates is declining i.e., $\dot{\hat{w}}_{R,t} - \dot{\hat{w}}_{H,t} < 0$ (see the next section), selection among college graduates decreases the average log ability among researchers if the marginal researcher has higher research-specific ability than the average researcher. Conditional on a given z_{Hi}, this case is illustrated by Case 1 in Panel (b) of Figure 1: as the return of working in research declines relative to the return of working in high-skill production, the research-specific ability of the college graduate on the margin between research and production declines, and thus the average research-specific ability of researchers also declines. Conversely, if the marginal researcher has lower research-specific ability than the average researcher, selection among college graduates increases the average research-specific ability among researchers; this scenario is depicted as Case 2 in Panel (b) of Figure 1. Note that this example is conditional on a given z_{Hi}; in general, we are averaging over all conditional distributions and thus, loosely speaking, each of the two cases illustrated in Figure 1 needs to hold on average.

2.4 Observed declines in research productivity: permanent or transitional?

What is a plausible range for the permanent diminishing returns to research? Rearranging the knowledge production function in equation (1) gives:

$$\Delta_k \ln\left(\frac{\dot{A}_t/A_t}{Z_{R,t}}\right) = (\phi - 1)\Delta_k \ln A_t, \tag{13}$$

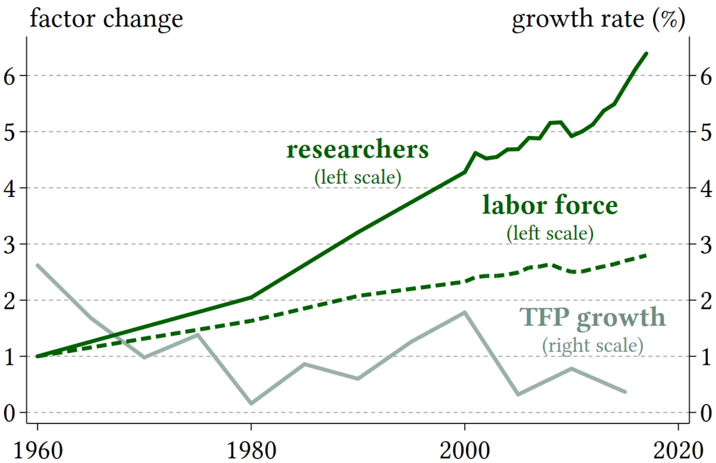
where Δ_k denotes the difference between $t + k$ and t and the equation holds for any $k > 0$. The left-hand side is the growth rate of research productivity; the equation thus just says that research productivity grows at a rate proportional to the growth rate of knowledge. The key here is that if $\phi < 1$, then research productivity is declining as knowledge grows. Supposing that we can observe some measure of the stock of ideas, we can directly use equation (13) to estimate ϕ. If the knowledge production function holds in the aggregate as assumed above, a natural measure of A is aggregate TFP; if it instead holds within a specific industry, a measure of A would be the citation-weighted stock of patents produced by that industry.

Absent a measure of the average ability of researchers, suppose we directly use the number of researchers to approximate $Z_{R,t}$. The resulting equation is then:

$$\Delta_k \ln\left(\frac{\dot{A}_t/A_t}{L_{R,t}}\right) = (\phi - 1)\Delta_k \ln A_t + \Delta_k \ln \overline{Z}_{R,t} \tag{14}$$

where $L_{R,t}$ is the number of researchers and $\overline{Z}_{R,t} = Z_{R,t}/L_{R,t}$ is the average human capital of researchers. Given the U.S. 1960-2017 time series of TFP (A_t) and researchers ($L_{R,t}$) in Figure 2, assuming no changes in researcher quality over time would give a $\phi = -0.8$. This is because the number of researchers has increased by a factor of over 6 over this time period; combining this observation with approximately constant TFP growth, the conclusion of strong diminishing returns is seemingly unavoidable. Note that this result is not unique to the aggregate level using

Figure 2: Factor changes in researchers vs. TFP growth, U.S. 1960-2017



Notes: The dark lines refer to the left vertical axis. The dark solid line shows the factor change in the number of researchers from 1960-2017, calculated by the authors as discussed in Section 3.2. The dark dashed line shows the factor change in the U.S. labor force from 1960, calculated from the Census/ACS. The light solid line refers to the right vertical axis, showing the average annual growth rate of TFP at a five-year frequency. The TFP growth rates are taken from the Private Business Sector multi-factor productivity growth series from the Bureau of Labor Statistics (2017). We match the scales of the two vertical axes so that proportional changes are comparable.

TFP growth as a measure of research output; Bloom et al. (2020) find low values of ϕ across a wide range of industries using various measures of research output.

However, there is no reason to suppose that a 6-fold expansion in the number of researchers could have been accomplished without changing the average ability of researchers. As shown above, if the marginal worker has low ability relative to the average researcher, the expansion of the research entails drawing lower-ability workers into the sector, thereby decreasing the average ability of researchers.

We refer to this composition effect as the "transitional" diminishing returns of research, in the sense that it vanishes when the economy reaches a long-run steady state with a constant share of researchers (see Appendix A.1 for balanced growth path characteristics). In the remainder of this paper, we will leverage the evolution of the researcher workforce and their earnings distribution, in conjunction with the model of researcher supply described above, to gauge the quantitative significance of our proposed mechanism. To be precise, we estimate the change in the average ability of researchers between 1960 and 2017, and then revisit the implied permanent diminishing returns using equation (14).

To clarify our point, note that said composition effect is different from typical diminishing returns to labor as described by knowledge production functions of the form $\dot{A}_t = A_t^{\phi} L_{R,t}^{\lambda}$, where $\lambda < 1$. The main difference is that the composition effect we focus on causes research returns to diminish with respect to the share of researchers in the labor force, whereas $\lambda < 1$ causes the returns to diminish with respect to the stock of researchers. As a result, the diminishing returns described by λ are isomorphic to those described by ϕ in determining long-run growth in semi-endogenous growth models and hence should be viewed as permanent.[12] To see this, consider a balanced-growth path with positive population growth and a constant share of researchers. In this case, there will be no composition effect as long as the ability distribution is stable regardless of the expanding population size. The marginal productivity of researchers in the case of $\lambda < 1$, however, would still be decreasing because the number of researchers continues to grow.

We also want to emphasize that the evolution of the earning distribution can distinguish between these two types of labor diminishing returns. For instance, the transitional diminishing returns change the dispersion of researchers earnings, whereas the permanent ones do not. Therefore, $\lambda < 1$ does not invalidate our approach because we infer the transitional composition effect using a Roy model estimated by individual-level data on earnings.

2.5 Relation to Bloom et al. (2020)

Bloom et al. (2020) propose correcting the time series of R&D expenditures for the average ability of researchers by dividing total nominal expenditures on intellectual property products by the average nominal earnings of college-educated male workers as a proxy for the average wage of reseachers. In our model, the average wage of researchers is given by

$$\overline{W}_{R,t} = W_{R,t} \times \mathbb{E}\left[e^{z_{Ri}+z_{Ci}} \mid \text{Researcher}\right], \tag{15}$$

with the expectation term denoting the average ability of researchers. Total expenditure on R&D is just $W_{R,t}L_{R,t}$, and so the measurement used by Bloom et al. would result in:

$$\hat{Z}_{R,t} = \frac{W_{R,t}Z_{R,t}}{W_{R,t} \times \mathbb{E}\left[e^{z_{Ri}+z_{Ci}} \mid \text{Researcher}\right]} = \frac{Z_{R,t}}{\mathbb{E}\left[e^{z_{Ri}+z_{Ci}} \mid \text{Researcher}\right]}. \tag{16}$$

The corrected series thus remains biased: in effect, the ideal deflator is the efficiency wage rate, but this wage rate cannot be deduced from the average wage without first knowing the average

[12]For instance, Jones (1995)'s model predicts the long-run growth rate of income per person to be $g_y = \lambda g_L/(1-\phi)$.

ability of researchers.

3 Empirical findings

The wage distribution is key to pinning down the self-selection in workers' ability (Heckman and Sedlacek, 1985, 1990; Heckman and Honoré, 1990). We exploit unique information on work activity from the National Survey of College Graduates (NSCG) to calculate among college graduates (i) the share of researchers, (ii) the relative earnings of researchers, and (iii) the earnings dispersion of researchers at the detailed occupation level. We then approximate aggregate trends in these moments by summing across occupations with the share of college graduates obtained from the Decennial Census and the American Community Survey (ACS) from 1960-2017. In Section 2, we will use the trends documented in this section to estimate self-selection in researchers' ability and the transitional diminishing returns.

The main challenge in measuring researchers' wages is constructing a researcher identifier; such information is not available in commonly used U.S. demographic data, such as the Decennial Census or the ACS. We exploit unique information on primary work activity from the NSCG to identify researchers. The NSCG is a biennial survey conducted by the National Science Foundation starting in 1993 which provides data on college graduates, with a focus on the science and engineering workforce.[13] The NSCG allows us to differentiate between researchers and non-researchers through reported work activity, but its short time series and focus on the college labor force limit its applicability to obtain the aggregate information of researchers.

To overcome this problem, we match the NSCG occupational codes into the Decennial Census for decades from 1960 to 2010 and the ACS for non-Census years between 2000 and 2017. We then calculate the number of researchers and moments of their earnings distribution by aggregating the corresponding occupational moments, weighted by the within-occupation researcher share as measured in the NSCG. The public-use Decennial Census sample includes 1 to 5% of the U.S. population (depending on the year), whereas the ACS samples 1% of the U.S. population, and thus working with these data sets significantly expands the sample of workers we have access to. In what follows we detail the specifics of our measurements, present results on the evolution of the researcher labor market, and briefly discuss some vulnerabilities of our analysis. We will use subscripts R to denote researchers, H to denote other college-educated workers (not including researchers), and N to denote workers without college degrees. We will denote all college

[13] In practice we are restricted to the public-use years of the NSCG: 1993, 2003, 2010, 2013, 2015, and 2017.

workers, including researchers, as $C = H \cap R$.

3.1 Defining researchers in the NSCG

To make progress, we must first take a stand on what workers should be counted as researchers. Through the lens of endogenous growth theory, a researcher is a worker who is involved in producing nonrival production inputs, e.g., new varieties, better-quality products, or more efficient processes (Romer, 1990). However, the extent to which any given output is non-rival is difficult to measure, and delineating exactly which workers enter the production function of such outputs is similarly challenging. To deal with this, the literature has generally adopted a purposefully-conservative definition of researchers by restricting to a subset of workers who are most clearly involved in nonrival production, namely scientists and engineers. For instance, (Jones, 1995, 2002, 2016) measures the number of researchers as "scientists and engineers engaged in R&D" as reported in the National Science Foundation's (NSF) innovation surveys.

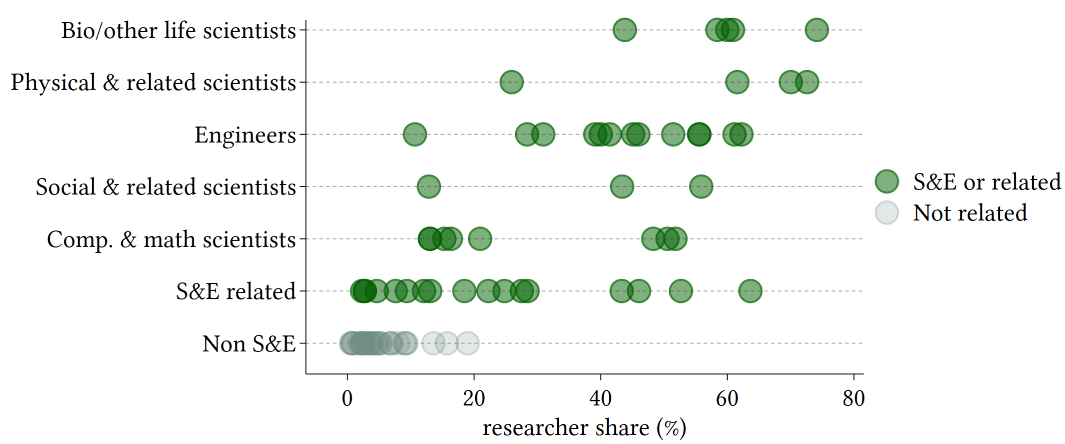
We hew to the spirit of these measurements while adapting and refining them to match our data source, which is a worker rather than a firm survey. Accordingly, we use self-reported primary work activity to identify researchers. Specifically, we label a worker as a researcher if their primary work activity (PWA) for their principal job is reported to be either "research" or "development." To preserve comparability to previous estimates in the literature, and to minimize the effect of noise on our measurement, we also restrict researchers to be in occupations categorized as "scientists and engineers" (S&E) or "S&E related" by the NSCG.[14] Formally, for every occupation j that we observe in the NSCG, let I_j be the set of workers in occupation j, and let L_j be the number of workers in I_j. We define the occupational researcher share, S_j^{NSCG}, as:

$$S_j^{\text{NSCG}} = \left[\frac{\sum_{i \in I_j} \mathbb{1}(\text{PWA}_i = \text{R\&D})}{L_j} \right] \times \mathbb{1}\left(j \in \left\{ \text{S\&E or related} \right\} \right), \tag{17}$$

where i indexes a worker and PWA_i denotes their primary work activity. Given that we use a broader definition of researchers, we expect to count more researchers than previous estimates relying on firm-reported surveys. We view this as an improvement because, as pointed out in Jones (2016) and Wolfe (2014), firm-reported R&D activities suffer from serious deficiencies. For example, only 30% of aggregate R&D is reported by non-manufacturing firms, and some companies with histories of rapid growth, such as Walmart, report no R&D at all.

[14] The former category includes occupations such as biochemist, mathematician, and computer scientist, whereas the latter includes various kinds of technicians.

Figure 3: Occupational researcher share in NSCG



Notes: Each circle represents an occupation in the NSCG. The vertical axis classifies occupations by major categories and the horizontal axis shows the share of workers within each occupation who report "research" or "development" as the primary work activity of their principal jobs. Data source: NSCG.

Figure 3 shows occupational researcher shares by broad occupational categories, pooling all publicly-available years of the NSCG (1993, 2003, 2010, 2013, 2015, 2017) and treating them as a single cross-section. The vertical axis reports seven major categories of occupations ranked in order of average researcher shares. The horizontal axis reports the share of respondents who report R&D as their primary work activity.

As expected, scientists and engineers have higher researcher shares than S&E-related and non-S&E occupations. Among scientists, occupations in the natural sciences (e.g. biochemists, physicists) have higher researcher shares than occupations in social, computer, and mathematical sciences. However, there is significant heterogeneity within each of the major occupational groups—even among S&E occupations, one-third have a researcher share smaller than 40%. These observations point out the limitations of measures solely based on occupations: not all scientists and engineers are researchers.

3.2 Trends in the U.S. researcher labor market

3.2.1 Share of researchers increases overall, decreases among college graduates

To measure long-term trends in the number of researchers and the first two moments of their log earnings, we impute the occupational-level share of researchers from the NSCG to the Decennial Census/ACS. We first match each occupation in the NSCG with one or more corresponding occupations in the Census/ACS.[15] For each year that we observe in the Census/ACS, we then measure the aggregate number of researchers as:[16]

$$L_R = \sum_{j \in J} L_{Cj} \cdot s_j^{NSCG}, \tag{18}$$

where L_{Cj} is the number of college graduates in occupation j, and J is the set of occupations in Census/ACS. Note that we restrict to the number of college graduates within each occupation because the NSCG only surveys college graduates, and thus s_j^{NSCG} measures the researcher share of employment among college graduates in occupation j. Although this limitation has the potential to bias our measurement, we believe that the number of researchers without college degrees is small enough over our sample period to avoid significantly compromising any of our conclusions.

Given the long sample period with which we work, there is a distinct possibility that the within-occupation researcher shares change over time. To address this concern, we calculate the average annual change rate of each NSCG occupation's researcher share and then impute fitted values into the Decennial Census/ACS assuming constant within-occupation change rates.[17] A major limitation in accounting for this concern is that the NSCG has relatively short coverage compared to the Decennial Census/ACS. To examine the robustness of our results to this extrapolation procedure, we consider two alternate ways of expanding the time series of researcher shares from the NSCG in Appendix C.

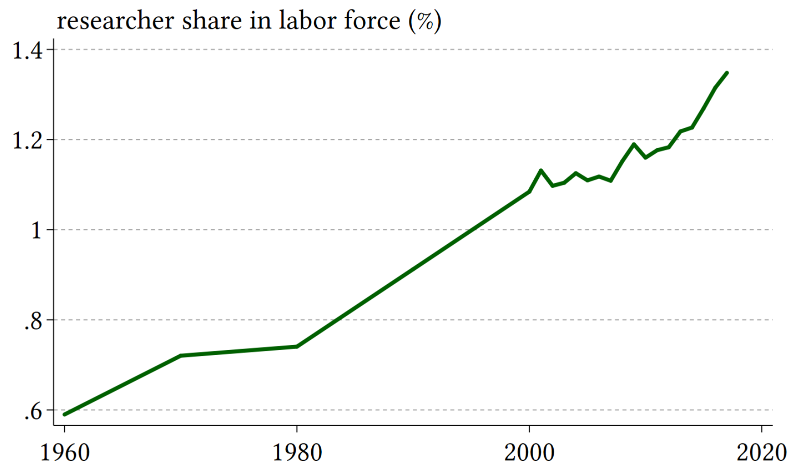
The resulting time series of the number of researchers is plotted in Figure 2. The dark solid line depicts the factor change in the number of researchers since 1960. As discussed in Section 2, our method of identifying researchers indicates that the U.S. number of researchers has increased 6-fold from 1960 to 2017 and that this increase is about two times more than that of the labor force. This implies that researchers as a share of the labor force have also increased. Figure 4 depicts the time series of the share of researchers in the labor force, $S_R = L_R/L$, over the period 1960-2017, where the labor force (L) is calculated from the Decennial Census and the ACS. The labor force share of researchers has increased significantly over time—from 0.6% in 1960 to 1.4% in 2017. The rising share of researchers motivates our focus on self-selection as a source of transitional diminishing returns of research: if researcher ability is positively self-selected, the expansion of

[15] The crosswalk is designed by the authors. The overlap between the NSCG and Census/ACS occupational codes is significant, but the Census/ACS codes are much finer for certain occupations that presumably do not employ many college graduates e.g. transportation workers. In practice, this is irrelevant to our results, since these occupations have 0 researcher shares.

[16] Variables are calculated using the Census/ACS unless otherwise specified.

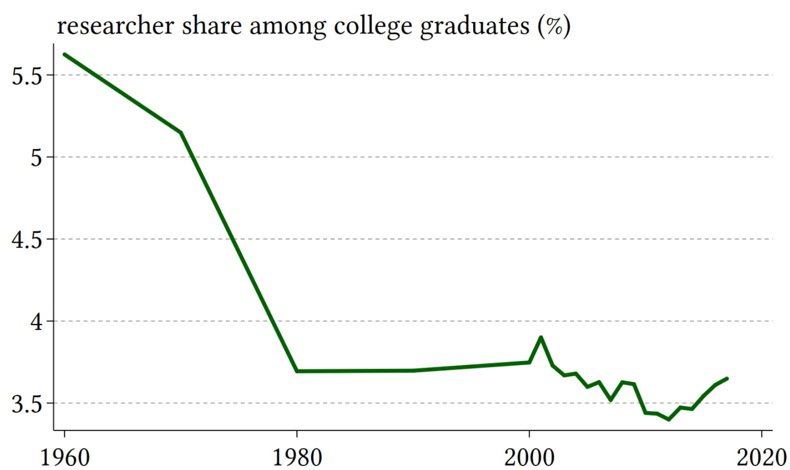
[17] To keep shares bounded between 0 and 1, we construct researcher share change rates from the change rates of the number of researchers and the number of non-researchers within each occupation.

Figure 4: Share of researchers in the labor force (S_R), U.S. 1960-2017



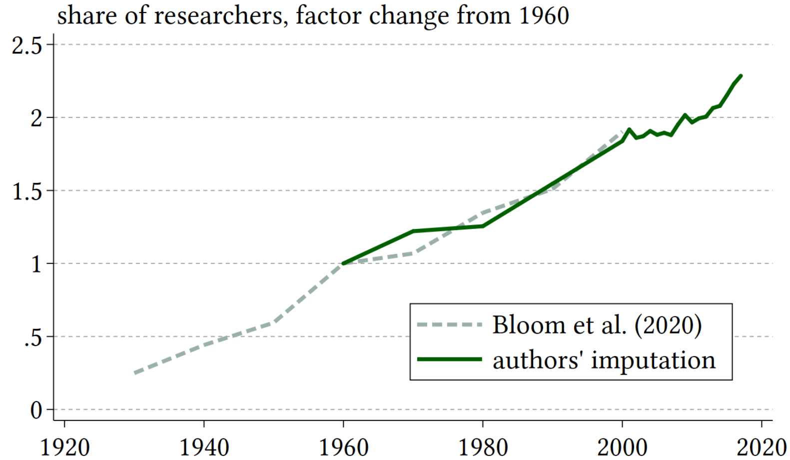
Notes: The number of researchers (L_R) is calculated by the authors using the NSCG and the Census/ACS. The total labor force (L) is calculated from the Census/ACS.

Figure 5: Share of researchers among college graduates ($S_{R|C}$), U.S. 1960-2017



Notes: The number of researchers (L_R) is calculated by the authors using the NSCG and the Census/ACS. The college-educated labor force (L_C) is calculated from the Census/ACS.

Figure 6: Share of researchers in labor force, authors' imputation vs. Bloom et al. (2020)



Notes: The solid line reflects the authors' calculation using the NSCG and the Census/ACS. The dashed line reflects Bloom et al. (2020)'s effective number of researchers divided by the U.S. labor force taken from FRED and the U.S. Bureau of Labor Statistics (BLS). Both series are normalized to reflect factor changes from 1960.

the research sector should lead to a decrease in average researcher ability.

In Figure 5, we further document the evolution of the researcher share among college graduates, $S_{R|C} = L_R/L_C$. We find that the share of researchers among college graduates has been declining over time, although the rate of decline slows down after 1980. The declining share of researchers among the college-educated labor force points to a potentially opposite force that changes the average ability of workers, depending on how researcher ability is self-selected among college workers. For example, if sectoral choices are positively sorted in ability, the declining trend in $S_{R|C}$ must have been causing the average ability of researchers to increase over time.

Before closing this subsection, we examine if our imputed trend is similar to other measures used in the literature. Figure 6 compares the factor change in the researcher share calculated in this paper with that in Bloom et al. (2020). The solid line depicts the authors' calculation using the NSCG and Census/ACS following equation (18). The dashed line, on the other hand, is calculated using Bloom et al. (2020)'s "effective number of researchers" divided by the U.S. labor force taken from FRED and the U.S. Bureau of Labor Statistics (BLS).[18] The figure shows that the growing

[18]Bloom et al. (2020) measure the effective number of researchers as the gross domestic investment in intellectual property products deflated by the nominal wage of males with a bachelor's or higher degree. The labor force for 1948-2000 is downloaded from the FRED website: https://fred.stlouisfed.org/series/CLF16OV, and the labor force for 1930 and 1940 are taken from *Population: Estimates of Labor Force, Employment, and Unemployment in the United*

share of researchers in our calculation is consistent with that in Bloom et al. (2020).

3.2.2 Researchers' earnings increasingly disperse over time

The rising share of researchers points to a potentially important role for transitional diminishing returns of R&D due to self-selection. As discussed in Section 2, average ability in an expanding sector decreases if abilities are positively sorted into the sector. A general feature of ability sorting is that the dispersion of sectoral earnings increases in the sectoral share of labor (Heckman and Honoré, 1990). To see whether this holds in the data, we measure the variance of researchers' log earnings, denoted as V_R, and compare it with the variance of log earnings for other workers.

To do this, we follow a similar logic to the previous subsection and exploit researcher's log earnings from the NSCG at the occupational level. Specifically, let e_i be the log earnings of worker i. For each occupation j we define:

$$S_{e,j}^{\text{NSCG}} = \left[\frac{\sum_{i \in I_j} e_i \cdot \mathbb{1}(\text{PWA}_i = \text{R\&D})}{\sum_{i \in I_j} e_i} \right] \times \mathbb{1} \left(j \in \left\{ \text{S\&E or related} \right\} \right). \tag{19}$$

We also calculate the occupational researcher shares of log earnings squared defined as:

$$S_{e2,j}^{\text{NSCG}} = \left[\frac{\sum_{i \in I_j} {e_i}^2 \cdot \mathbb{1}(\text{PWA}_i = \text{R\&D})}{\sum_{i \in I_j} {e_i}^2} \right] \times \mathbb{1} \left(j \in \left\{ \text{S\&E or related} \right\} \right). \tag{20}$$

We can then measure the variance of log earnings of researchers by summing across occupations with weights obtained from the Census/ACS.

$$V_R = \frac{1}{L_R} \left[\sum_{j \in J} \left(\sum_{i \in C_j} {e_i}^2 \right) s_{e2,j}^{NSCG} \right] - \tilde{E}_R^2, \tag{21}$$

where C_j is the set college-educated workers in occupation j obtained from Census/ACS. The term $\tilde{E}_R$ denotes the average log earnings of researchers, calculated by

$$\tilde{E}_R = \frac{1}{L_R} \left[\sum_{J} \left(\sum_{i \in C_j} e_i \right) s_{e,j}^{NSCG} \right]. \tag{22}$$

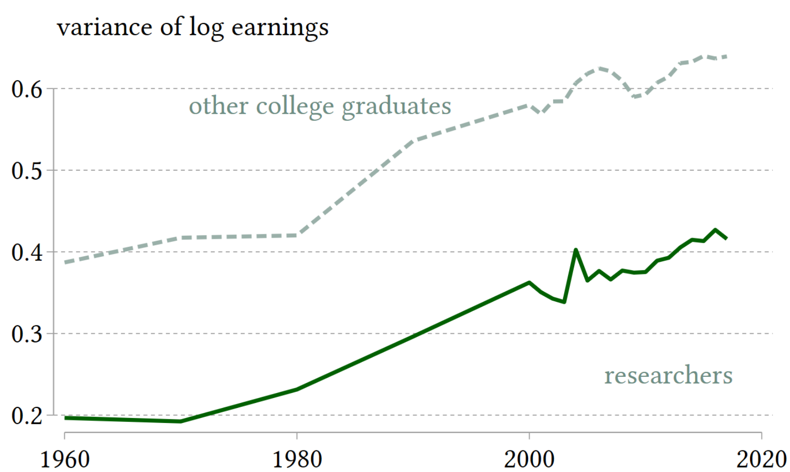
We use the same approach to obtain the mean and variance of log earnings for other college

States, 1940 and 1930 by Bureau of the Census, printed by U.S. Government Printing Office, Washington, 1944.

Figure 7: Average log earnings, relative to no-college workers, U.S. 1960-2017.

Notes: Figure depicts the log relative earnings relative to workers without college degrees. The solid line depicts the series for researchers ($E_R = \tilde{E}_R - \tilde{E}_N$) and the dashed line depicts that for other college graduates ($E_H = \tilde{E}_H - \tilde{E}_N$). Source: authors' calculation from the NSCG and the Census/ACS.

Figure 8: Variance of log earnings, U.S. 1960-2017.



Notes: Figure depicts the variance of log earnings The solid line depicts the series for researchers (V_R) and the dashed line depicts that for other college graduates (V_H). Source: authors' calculation from the NSCG and the Census/ACS.

graduates. Specifically,

$$\tilde{E}_H = \frac{1}{L_H}\left[\sum_{J}\left(\sum_{i\in C_j} e_i\right)\left(1 - s^{NSCG}_{e,j}\right)\right],\quad V_H = \frac{1}{L_H}\left[\sum_{j\in J}\left(\sum_{i\in C_j} {e_i}^2\right)\left(1 - s^{NSCG}_{e2,j}\right)\right] - \tilde{E}^2_H. \tag{23}$$

The mean and variance of log earnings for no-college workers, $\tilde{E}_N$ and V_N are then directly calculated from the ACS.[19]

Figure 7 reports the imputed mean log relative earnings for researchers $E_R = \tilde{E}_R - \tilde{E}_N$ and for other college-educated workers $E_H = \tilde{E}_H - \tilde{E}_N$. Both college-educated sectors exhibit higher average earnings than the no-college sector, with the research sector exhibiting the highest average earnings. Figure 8 reports the imputed variance of log earnings V_R and V_H. The results are consistent with the prediction of a standard Roy model with positive self-selection in ability: the dispersion of earnings has been increasing in expanding sectors, i.e., researchers and other college-educated workers. We report more details about the aggregate earning trends in Appendix E. First, we show that while the variance of log earnings in the no-college sector (V_N) also increases over time, the magnitude of the increase is smaller than that in the other two sectors. Second, we show that the trends we document are robust to residualizing log earnings in the ACS, controlling for age, age-squared, gender, race, and year-fixed effects.

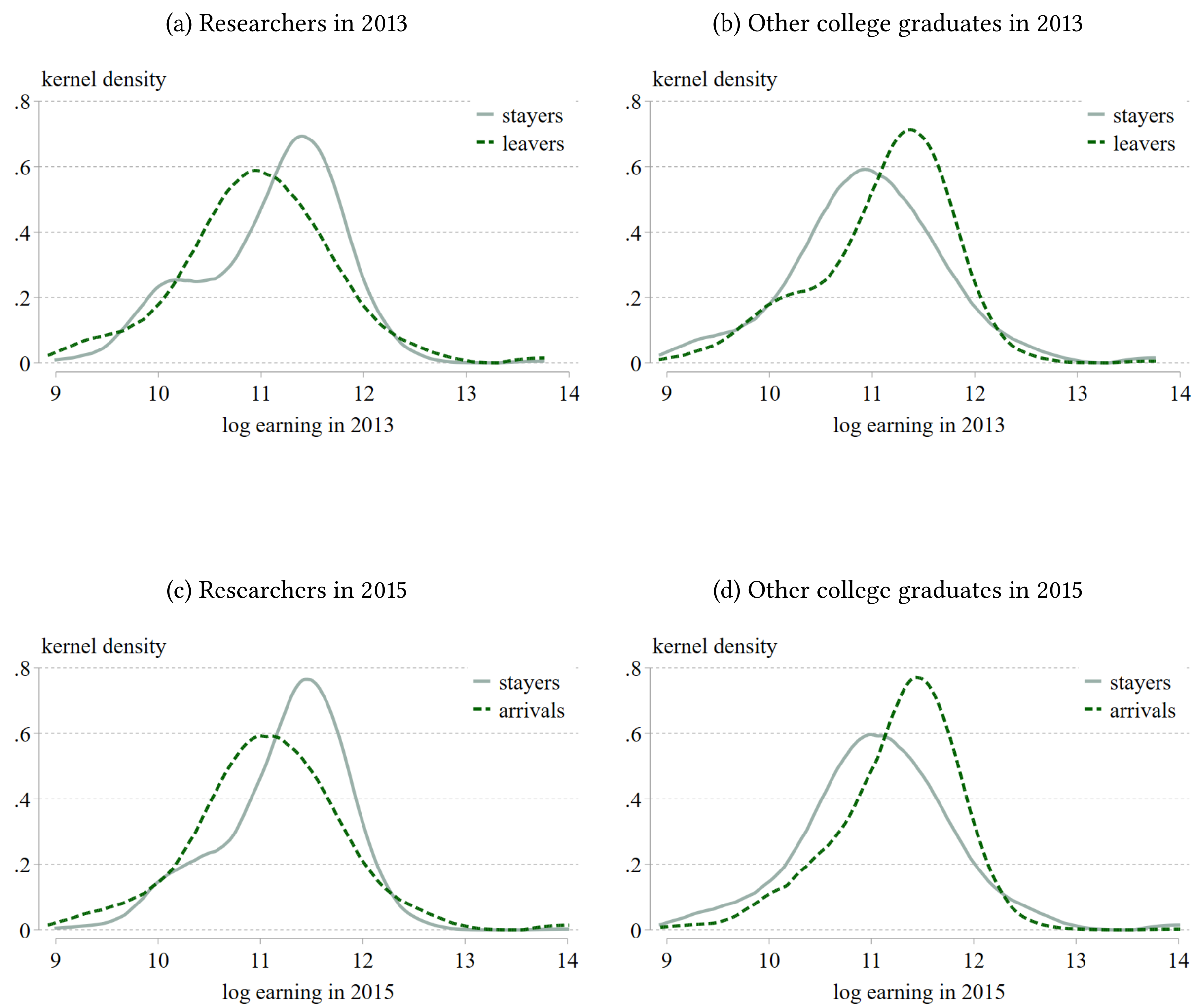
3.3 Evidence for self-selection from longitudinal data

Cross-sectional earnings data are not enough to identify a Roy model without using higher-order moments such as skewness; for any Roy model with correlated abilities, there exists a Roy model with independent abilities that leave the first two moments unchanged (Heckman and Honoré, 1990). Therefore, in this subsection, we make use of a short panel constructed from the NSCG to examine the relative earnings of workers who transition between the research (R) and non-research (H) sectors among college graduates. Assuming that workers with weaker comparative advantage in a sector are more likely to transition between sectors, comparing the mean log earnings of workers who move between sectors to those of workers who stay in a sector allows us to compare the average worker in a sector to the marginal worker in that sector.

The NSCG applies time-consistent respondent identifiers in the 2010, 2013, and 2015 surveys.

[19] We could in principle use the occupational researcher employment shares to impute the first two moments of researchers' log earnings, though we prefer our measures because they are less biased. However, in practice this choice makes little difference: Appendix D compares the different occupational researcher shares used in the imputation of these moments and finds that they are almost identical.

Figure 9: Earning distributions in the NSCG 2013-2015 panel data



Notes: Figure depicts the log earning distributions for sectoral stayers and movers in the NSCG 2013-2015 panel. Panels (a) and (c) show those of researchers in 2013 and 2015 respectively; Panels (b) and (d) show those of non-researcher college graduates in 2013 and 2015 respectively. Data source: NSCG 2013-2015 panel.

Therefore, a subsample of respondents is included in multiple of these three survey years. In our sample, 25,621 respondents are observed in both 2010 and 2013, and 36,889 respondents are observed in both 2013 and 2015. We refer to the former as the NSCG 2010-2013 panel and the latter as the 2013-2015 panel. For those respondents, we can observe whether they move across occupations and their earnings conditional on doing so. For the remainder of this subsection, we focus on the 2013-2015 panel because of its larger sample size. In Appendix F, we show that all findings in this subsection are robust in the 2010-2013 panel. Moreover, we take the average of the moments over the 2010-2013 and 2013-2015 panels when using the longitudinal moments to estimate the model in Section 2.

To implement the aforementioned comparison, we define movers as workers who experienced both a change in primary work activity (between R&D and non-R&D) and a change in occupational code between 2013 and 2015.[20] In the resulting sample, 23.8% of researchers in 2013 move out of research in 2015; 3.6% of non-researcher college workers move into research in 2015. We label these workers as sector "leavers" in the 2013 sample. On the other hand, 20.5% of researchers in 2015 were not researchers in 2013; 4.3% of non-researchers in 2015 were researchers in 2013. We label these workers as sector "arrivals" in the 2015 sample. We then compare the distribution of log earnings of movers to that of stayers in both the origin and destination sectors. Figure 9 implements this comparison. The top two panels compare leavers' to stayers' earnings in the origin sector in 2013. The bottom two panels perform the same comparison between stayers and arrivals in the destination sector in 2015. Finally, the left two panels report the earning distributions in the research sector (R), and the right two panels report those in the non-research college sector (H).

The figure reveals a striking pattern: movers in the non-research college sector have systematically higher earnings than stayers, whereas the opposite is true in the research sector. On average, stayers earn about 0.06 log points more than movers (both leavers and arrivals) in the research sector; but movers earn 0.16 log points more than stayers in the non-research sector. As we will argue in Section 2, this can only happen in a standard Roy model if the marginal worker in the non-research college sector has higher ability than the average worker, usually referred to as negative self-selection. Conversely, the lower earnings of movers relative to stayers in the research sector suggest that the marginal worker in the research sector has lower ability than the average worker, i.e., positive self-selection.

Appendix F reports more details about the earnings distributions in the NSCG panel. Figure

[20] The latter restriction is made to avoid the effects of measurement error in workers' primary work activity.

A7 shows the analogous moments of 9 in the 2010-2013 panel. Table A1 reports the mean, standard deviation, and number of observations for each distribution drawn in the two figures. We also show in Figure A8 that the primary findings in this subsection are robust to using residual log earnings controlling for age, age-squared, gender, race, and year-fixed effects.

3.4 Beyond the United States

Before proceeding to the model, we want to highlight that the facts documented in Section 3.2 are valid in other developed countries beyond the United States, suggesting that the forces we highlight are not unique to the U.S. context. In this section, we present the cross-time and cross-country patterns of R&D workforces for 18 countries whose time series of annual R&D personnel from 1980 to 2010 are available in OECD Statistics.[21] We calculate the total number of researchers, the population share of researchers, and the share of researchers among college graduates for each country and each five-year interval from 1980 to 2010. The number of researchers is taken from OECD Statistics and is defined as the full-time equivalent number of persons engaged in R&D.[22] Data for college graduates are taken from Barro and Lee (2013), and is defined as the number of workers with a complete tertiary education. Data for GDP per person are taken from the Penn World Table (PWT) 9.1 (Feenstra, Inklaar, and Timmer, 2015). Data for TFP growth rates are also from OECD Statistics.

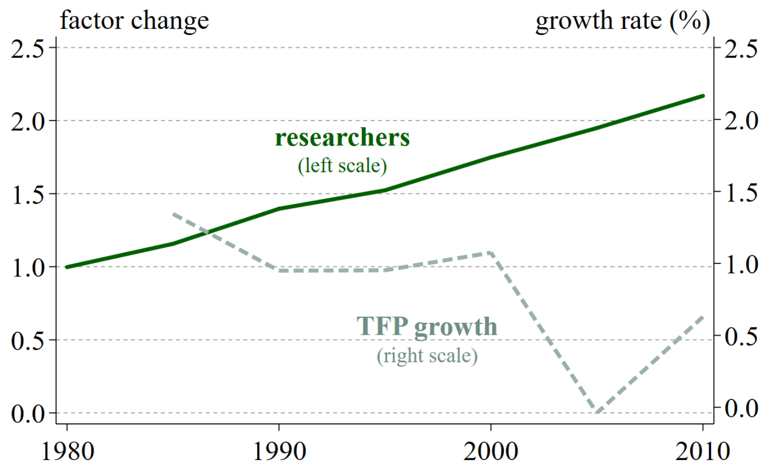
We first show that average researcher productivity has also been declining over time for the 18 OECD countries. To see this, we take the population-weighted average over the 18 countries and examine the time trends of average annual TFP growth and the factor change in the number of researchers since 1980. The results are presented in Figure 10. While the number of researchers doubled from 1980 to 2010, the average annual growth rate of TFP was flat at around 1% before 2000 and dropped during the 2003-2007 interval. This implies that the average productivity of researchers has fallen to less than half of its 1980 level.

We next examine the trend of average researcher-to-population shares for the 18 OECD countries. Panel (a) of Figure 11 shows the population-weighted average share of researchers (S_R) over time, and Panel (b) shows the researcher shares relative to the college-educated population

[21]The 18 countries are: Australia, Austria, Belgium, Canada, Denmark, France, Germany, Hungary, Iceland, Ireland, Italy, Japan, Netherlands, Norway, Spain, Sweden, United Kingdom, and the United States. Data downloaded from the *OECD Statistics* website: https://stats.oecd.org/.

[22]The methodology for measuring R&D personnel in this data set is documented in the OECD's Frascati Manual: http://www.oecd.org/sti/inno/Frascati-Manual.htm.

Figure 10: Factor changes in researchers vs. TFP growth, average over 18 OECD countries



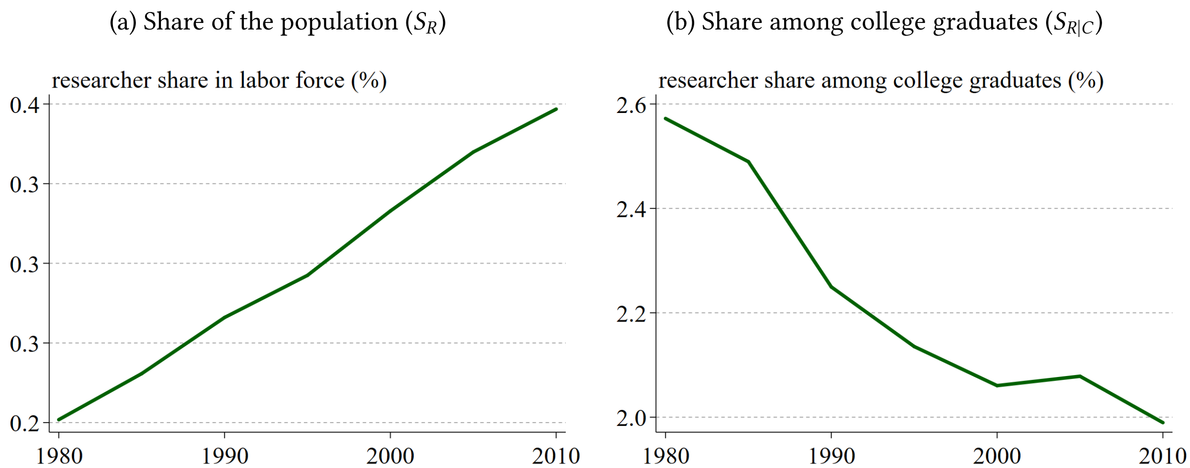
Notes: The solid line refers to the left vertical axis, depicting the factor change of researchers since 1980. The dashed line refers to the right vertical axis, depicting the average annual growth rate of TFP. Data are reported in five-year intervals. We match the scales of the two vertical axes so that proportional changes of the two series are comparable. Data source: see Section 3.4.

($S_{R|C}$).[23] The direction of both trends is consistent with their U.S. counterparts as presented in the previous section. The overall population share of researchers doubled from 1980 to 2010, and the share among college graduates in 2010 is around four-fifths of its 1980 level.

Figure 12 shows the cross-country analog for the evolution of researcher shares along the path of development. Panel (a) plots the overall share of researchers (S_R) by log GDP per person for each country-year observation. It shows a clear positive correlation between researcher share and per-person income. We further project the researcher shares on log income per person controlling for year-fixed effects (regressions are weighted by population size). The estimated slope is 0.18 with a standard error of 0.03, meaning that, on average, a country with a 1 percentage point higher per capita income has a 0.18 percentage point higher researcher share. Panel (b) plots the researcher share among college graduates ($S_{R|C}$) by per capita log GDP, showing an analogous pattern to panel (b) of Figure 11. Projecting the shares on log GDP per person controlling for year fixed effects, we find that a country with a 1% higher per capita income has, on average, a 1.6 percentage point lower $S_{R|C}$, with standard error 0.55. We conclude that the cross-time and cross-country patterns are both consistent with those observed in the United States.

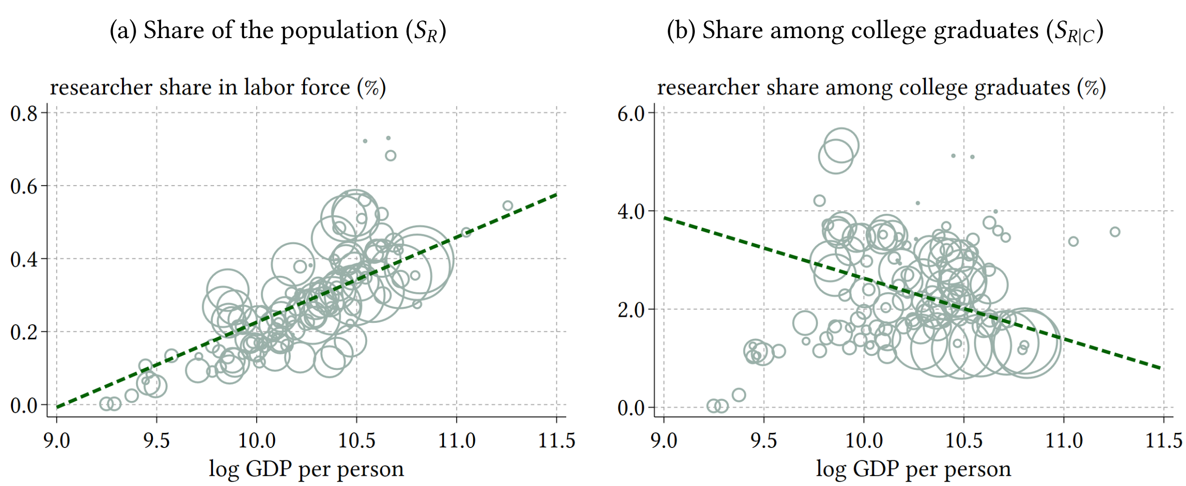
[23]Note that here we calculate the population share, instead of the labor force share of researchers as in the previous subsection. This is because Barro and Lee (2013) only provides the population shares for education groups. To further refine our measurement, we may replace it with data from the International Labor Organisation (ILO) to obtain employment data by education level. This way, we will be able to measure the employment share of researchers.

Figure 11: Time trend of researcher shares, average over 19 OECD countries



Notes: Panel (a) depicts the share of researchers relative to the total population (S_R). Panel (b) depicts the share of researchers relative to college graduates ($S_{R|C}$). Data are reported in five-year intervals. Data source: see Section 3.4.

Figure 12: Researcher shares vs. income per person, 19 OECD countries, 1980-2010



Notes: Each circle depicts the share of researchers by its log income per person for a country-year observation (18 countries, each has 7 observations from 1980 to 2010 in five-year frequency). The sizes of the circles represent countries' relative populations. The dashed lines depict the fitted value of population-weighted OLS regression. Panel (a) reports the overall share of researchers, while panel (b) reports the share of researchers among college graduates. Data source: see Section 3.4.

4 Estimation

In this section, we return to the model described in Section 2 and use it to estimate the change in the average ability of researchers between 1960 and 2017. We estimate the model using moments of the cross-section of researchers' earnings distribution and thus drop time subscripts from hereon.[24] We parameterize the model flexibly enough to replicate the patterns documented in Sections 3.2 and 3.3. We then use said patterns to estimate the model using indirect inference and employ the estimated model to infer the change in average researcher ability over our sample period. Finally, we discuss how our findings change the implied magnitude of the permanent diminishing returns of research, i.e., ϕ as defined in Section 2.

4.1 Parameterization

We start with parameterizing our model. Recall that our goal is to pin down the changes in average ability of researchers over time. We do not need to identify every single structural parameter in the model, but a subset of them that are enough to identify researchers' average ability. Specifically, we do not need to separately identify μ_C and $w_N + \mu_N$. Therefore, we will assume $w_N + \mu_N = 0$, and hence the implied μ_C combines all forms of opportunity costs of going to college.

Let s denote an indicator for going to college, the educational choice problem becomes:

$$u_i = \max_{s \in \{0,1\}} \{s \cdot (z_{Ci} + u_{Ci} - \mu_C)\}, \tag{24}$$

where u_{Ci} is the is the expected per-efficiency-unit utility gain from realizing sector-specific ability as defined below:

$$u_{Ci} = \mathbb{E}\left(\max_{j \in \{R,H\}} \{z_j + w_j + \mu_j\} \,\middle|\, z_{Ci} \right). \tag{25}$$

From here on, we will refer to $s \in \{0, 1\}$ as the solution to the educational choice problem and $j \in \{H, R\}$ as that for the sectoral choice problem; these are implicitly functions of worker types and sectoral returns, but we suppress this dependency for brevity's sake.

We parameterize the ability distribution with a joint log normal form. The sector-neutral

[24] We discuss how we construct model counterparts to the moments relating to movers/stayers in Section 4.3.

return to college, z_C, is drawn from a log-normal distribution with zero mean and variance σ_C^2. We assume that the sector-neutral ability is independent of sector-specific abilities. Conditional on attending college, workers then draw sector-specific abilities (z_H, z_R) from a bivariate log-normal distribution with zero means, variances σ_R^2 and σ_H^2, and correlation coefficient ρ.[25]

Under these assumptions, we can solve the model backward. The independence between sector-neutral and sector-specific abilities implies that sectoral choices only depend on the joint distribution of z_{Ri} and z_{Hi}. The log-normal form allows us to solve the sectoral shares analytically. Let $\tilde{w}_j = w_j + \mu_j$ for $j \in \{H, R\}$ and $\theta = \sqrt{\sigma_R^2 + \sigma_H^2 - 2\rho\sigma_R\sigma_H}$. The share of researchers conditional on attending college is:

$$S_{R|C} = \Pr(j = R \mid s = 1) = \Phi\left(\frac{\tilde{w}_R - \tilde{w}_H}{\theta}\right) \tag{26}$$

where $\Phi(\cdot)$ denotes the cumulative density function (CDF) of standard normal distribution. The independence assumption also implies that all workers have the same expected gain of sector-specific ability before going to college, namely, $U_C(z_{Ci}) = U_C$ for all z_{Ci}. Let

$$u_C \equiv \mathbb{E}\left(\max_{j \in \{R,H\}} \{z_{ji} + \tilde{w}_j\} \,\middle|\, z_{Ci}\right), \tag{27}$$

then the log-normal form allows us to rewrite equation (27) as:

$$u_C = \tilde{w}_R \Phi\left(\frac{\tilde{w}_R - \tilde{w}_H}{\theta}\right) + \tilde{w}_H \Phi\left(\frac{\tilde{w}_H - \tilde{w}_R}{\theta}\right) + \theta\phi\left(\frac{\tilde{w}_R - \tilde{w}_H}{\theta}\right), \tag{28}$$

where $\phi(\cdot)$ denotes the probability density function (PDF) of a standard normal distribution. Variation in educational choices is thus solely determined by heterogeneity in z_C. The share of college-educated workers is then:

$$S_C = \Pr(s = 1) = \Phi\left(\frac{u_C - \mu_C}{\sigma_C}\right). \tag{29}$$

The unconditional share of researchers is thus $S_R = S_C \times S_{R|C}$ and the share of other college graduates is $S_H = S_C \times \left(1 - S_{R|C}\right)$.

There are two main sources of self-selection in this model: one from educational choices

[25] We cannot separately identify the means of the ability distributions and sectoral efficiency wage rates and thus normalize the mean of ability for all sectors to be zero. This means that we only aim to quantify the compositional part of the aggregate change in average ability that is due to self-selection.

and the other from sectoral choices. Under the independence assumption, the distribution of z_C governs the self-selection in college education, whereas the joint distribution of (z_R, z_H) governs the self-selection in sectoral choices. To see this, express average log earnings of each sector $j \in \{H, R\}$ as:

$$E_j = w_j + \underbrace{\left(\frac{\sigma_C}{S_C}\right)\phi\left(\frac{u_C - \mu_C}{\sigma_C}\right)}_{\mathbb{E}(\ln z_C \mid s=1)} + \underbrace{\left(\frac{\sigma_j^2 - \rho\sigma_j\sigma_k}{\theta\, S_{j|C}}\right)\phi\left(\frac{\tilde{w}_j - \tilde{w}_k}{\theta}\right)}_{\mathbb{E}(\ln z_j \mid s=1, j=j)}, \tag{30}$$

where $k = \{R, H\} \setminus j$.

Equation (30) points to the potential of estimating the model using the sectoral earning distributions. In the spirit of Heckman and Honoré (1990), we incorporate the dispersion of log earnings for each sector: $V_j = \text{Var}(e_R | s = 1, j = j)$ for each $j \in \{R, H\}$ in addition to the sectoral mean. As described in Section 3.3, we also make use of the panel moments, i.e., those comparing sectoral movers' and stayers' earnings from 2013-2015 NSCG, to help identify the joint distribution of sector-specific abilities. While we do not have analytical solutions for these moments, we will calculate them by simulating the model for the purposes of estimation.[26]

4.2 Discussion of distributional assumptions

There are two substantive restrictions made on the joint distribution of ability: that sector-neutral high-skill ability is independent of its sector-specific counterparts, and that the joint distribution takes a log-normal form. We discuss each in turn.

First, assuming that sector-neutral ability is independent of sector-specific counterpart is equivalent to assuming that self-selection due to educational choice is positive, and hence the ability distribution only governs the magnitude of this positive selection. We are comfortable with this simplification because positive selection into education is well documented in the literature (see Willis and Rosen, 1979; Card, 1999; Carneiro and Lee, 2011; Heckman, Humphries, and Veramendi, 2018, among others). The direction and magnitude of self-selection in sectoral choices, on the other hand, is flexibly governed by the joint distribution of sector-specific abilities.

The main alternative to the log-normal assumption when studying the efficiency implica-

[26] Note that our model has no between-sector flows. To generate counterparts for the data longitudinal moments, we rank simulated observations by their comparative advantage in a sector and define movers as having comparative advantage below a given percentile, where we match this percentile to the mover share in the data.

tions of labor reallocation using Roy-type models is to assume a max-stable form, such as bivariate Fréchet or bivariate Pareto (e.g. Lagakos and Waugh, 2013; Hsieh et al., 2019). Although these distributions significantly simplify the model by providing analytical solutions to the sector-choice problem, they restrict sectoral average earnings to be invariant with respect to changes in wage rates. This means that the direction and magnitude of self-selection are fixed among this class of distributions so that a 1% increase in the efficiency wage rate of a sector always results in a 1% drop in its average ability. Our goal, in contrast, is to quantify the direction and magnitude of self-selection with data on earnings, and so such a parameterization is too restrictive in our case.

4.3 Estimation by indirect inference

We estimate the model by indirect inference. The set of estimated parameters is listed in Table 1 and denoted as vector $\boldsymbol{p}$. As we will show below, our targeted data moments include initial labor market outcomes and their changes over the period 1960-2017. To generate these moments in the model, we assume that the ability distribution does not change over time; on the other hand, changes in the relative sectoral wages and costs drive changes in labor market outcomes between 1960 and 2017. Therefore, we categorize the parameters into three groups in the table. The first group governs the latent ability distribution, which is invariant over time. The second group is the sectoral wages and costs in 1960.[27] The third group is the change in relative wages and costs from 1960 to 2017. Note that, to pin down the changes in average researcher ability, we do not need to separately identify all structural parameters, but the transformation of parameters as listed in the table.

The set of targeted data moments is listed in Table 2. We partition the moments into two categories: the exactly-matched moments (dented as $\boldsymbol{D}_1$) and the numerically-approached moments (denoted as $\boldsymbol{D}_2$). Specifically, we define the indirect inference estimate as:

$$\hat{\boldsymbol{p}} = \arg\min_{\boldsymbol{p}} \left[\boldsymbol{M}_2(\boldsymbol{p}) - \boldsymbol{D}_2\right]' \boldsymbol{W} \left[\boldsymbol{M}_2(\boldsymbol{p}) - \boldsymbol{D}_2\right], \quad \text{subject to} \quad \boldsymbol{M}_1(p) = \boldsymbol{D}_1, \tag{31}$$

where $\boldsymbol{W}$ is a weighting matrix.[28] As listed in Panel (a) of the table, we exactly match the initial sectoral shares and their changes over the period 1960-2017. This enables us to make use of

[27] Note that we can only identify two of three sectoral cost shifters; this follows from the expressions for the sectoral shares, i.e., equations (26), and (29), and sectoral mean log earnings, equation (30): realized sectoral shares only identify relative sectoral returns. Whereas mean log earnings allow for separate identification of the sectoral efficiency wage rates, no similar moment is available for separately identifying the sectoral preference shifters.

[28] In practice we only use the weighting matrix to put the distance between model and data moments in the same units, namely percent deviation from the data moments.

Table 1: Estimated parameters

Parameter	Description	Value
	Latent ability distribution	
σ_C	variance of ability gained from college	1.335
σ_R	variance of ability in sector R	0.258
σ_H	variance of ability in sector H	0.210
ρ	correlation between abilities in sector R and H	0.961
	Sectoral wages and costs in 1960	
w_R	log efficiency wage of sector R	-1.869
w_H	log efficiency wage of sector H	-1.928
μ_C	log costs of college education	-0.252
$\mu_H - \mu_R$	log costs of becoming a researcher	0.187
	Changes in relative wages and costs from 1960 to 2017	
$\Delta(\tilde{w}_R - \tilde{w}_H)$	relative log wage-to-cost ratio	-0.144
$\Delta(u_C - \mu_C)$	mean net log return of college	0.445

Table 2: Targeted moments, data vs. model

Panel (a): Exactly-matched moments, $\boldsymbol{D}_1 = \boldsymbol{M}_1(\hat{\boldsymbol{p}})$

Moment	**Notation**	**Data**
Initial sectoral shares in 1960		
share of college grads (%)	S_C	10.49
share of researchers among college grads (%)	$S_{R\|C}$	5.63
Changes in sectoral shares from 1960 to 2017		
share of college grads (pp)	ΔS_C	26.5
share of researchers among college grads (pp)	$\Delta S_{R\|C}$	-1.98
Longitudinal moments in 2010-2015 NSCG		
share of movers in R (%)	$S_{R\to H}$	24.0
share of movers in H (%)	$S_{H\to R}$	3.5

Panel (b): Numerically-approached moments, $\boldsymbol{D}_2$ vs. $\boldsymbol{M}_2(\hat{\boldsymbol{p}})$

Moment	**Notation**	**Data**	**Model**
Initial mean log earnings in 1960			
relative mean log earnings of researchers	E_R	0.801	0.801
relative mean log earnings of other college grads	E_H	0.372	0.372
Changes in earnings dispersion from 1960 to 2017			
researchers	ΔV_R	0.212	0.223
other college grads	ΔV_H	0.243	0.224
Longitudinal moments in 2010-2015 NSCG			
mean log earnings, leavers minus stayers in R	$E_R^{R\to H}$	-0.080	-0.079
mean log earnings, leavers minus stayers in H	$E_H^{H\to R}$	0.180	0.172

the analytical solutions for sectoral shares, i.e., equations (26) and (29), as restrictions on the parameters that effectively reduce the dimensionality of the parameter space in the algorithm. The shares of sectoral movers in the 2010-2015 NSCG panel are used when generating earnings for movers vs. stayers in the indirect inference, which will be illustrated below.

Panel (b) of the table listed the numerically-approached moments ($\boldsymbol{D}_2$). We categorize them into three groups. The first group includes mean log earnings among researchers in 1960 (E_R) and among non-researcher college graduates in 1960 (E_H). The second group includes the change in the variance of log earnings among researchers over the period 1960-2017, i.e., ΔV_R as defined in Section 3.2, and the analogous moment for non-researcher college graduates, ΔV_H. The third group is the longitudinal moments as documented in Section 3.3. This includes $E_R^{R \to H}$, the mean log earnings of researchers moving out of the research sector minus those of stayers, averaged over the two panels 2010-2013 and 2013-2015; and $E_H^{H \to R}$, the analogous longitudinal moment for non-researcher college graduates.

We construct model counterparts for these moments through a combination of analytical solutions and simulation. Because the random variable $z_C \mid s = 1$ is a truncated normal, its first two moments are known functions of parameters. Moreover, the distribution of $z_C \mid s = 1$ does not depend on whether an individual chooses to become a researcher or not. As a result, we do not need to draw z_C in our simulations. Specifically, given a simulation of z_R and z_H, we can calculate mean log earnings by

$$E_j = w_j + \underbrace{\mathbb{E}\left(z_{Ci} \mid s = 1\right)}_{\text{analytical}} + \underbrace{\mathbb{E}\left(z_{ji} \mid s = 1, j = j\right)}_{\text{simulated}}, \tag{32}$$

for each $j \in \{R, H\}$. Likewise, we construct model counterparts for ΔV_R and ΔV_H by

$$\Delta V_j = \underbrace{\Delta \text{Var}\left(z_{Ci} \mid s = 1\right)}_{\text{analytical}} + \underbrace{\Delta \text{Var}\left(z_{ji} \mid s = 1, j = j\right)}_{\text{simulated}}, \tag{33}$$

for each $j \in \{R, H\}$. To construct model counterparts for longitudinal moments, $E_R^{R \to H}$ and $E_H^{H \to R}$, we first determine a set of sectoral movers in our model. To do so, we assume that movers are those on the margin of choosing the other sector, i.e., those with weaker comparative advantage in their sector. To be precise, we pick the thresholds, $\overline{\alpha} > \underline{\alpha}$, such that the sectoral shares of movers in the 2010-2015 NSCG panel are exactly replicated:

$$S_{R\to H} = \Pr\left(\tilde{w}_R - \tilde{w}_H < z_R - z_H < \overline{\alpha}\right) \tag{34}$$

$$S_{H\to R} = \Pr\left(\underline{\alpha} < z_{Ri} - z_{Hi} < \tilde{w}_R - \tilde{w}_H\right), \tag{35}$$

as listed in Panel (a) of Table 2. Given the thresholds, the mean log earnings of movers relative to stayers in the model are then calculated by simulation:

$$E_R^{R\to H} = \mathbb{E}\left(e_{Ri} \mid \tilde{w}_R - \tilde{w}_H < z_{Ri} - z_{Hi} < \overline{\alpha}\right) - \mathbb{E}\left(e_{Ri} \mid z_{Ri} - z_{Hi} > \overline{\alpha}\right) \tag{36}$$

$$E_H^{H\to R} = \mathbb{E}\left(e_{Hi} \mid \underline{\alpha} < z_{Ri} - z_{Hi} < \tilde{w}_R - \tilde{w}_H\right) - \mathbb{E}\left(e_{Hi} \mid z_{Ri} - z_{Hi} < \underline{\alpha}\right), \tag{37}$$

Summing up, our estimation algorithm consists of the following steps. First, take a candidate minimizer $\boldsymbol{p}$ such that the moments in Panel (a) of Table 2 are exactly matched, i.e., $\boldsymbol{M}_1(p) = \boldsymbol{D}_1$. Second, simulate the model and calculate the model counterpart of moments listed in Panel (b) of Table 2, i.e., $\boldsymbol{M}_2(p)$. Third, numerically search for the estimates for parameters, $\hat{\boldsymbol{p}}$, to numerically approach D_2 with $\boldsymbol{M}_2(p)$ as defined in equation (31).

4.4 Estimation results

Table 1 reports the estimated parameters, and Table 2 compares the targeted moments in the data and model. The model is able to match the targeted moments quite closely. In this subsection, we illustrate how the moments help quantify the two channels of self-selection.

Self-selection in education is mainly governed by how the earning dispersion among college graduates changes with college expansion, i.e., ΔV_R and ΔV_H. Recall that education is positively self-selected by design. College expansion should result in a larger increase in earning dispersion among college graduates if there is a larger gap between the marginally selected and other college graduates, which means a stronger self-selection. Another way of seeing this is to express the average log ability conditional on college graduates as follows:

$$\mathbb{E}\left(z_{Ci} \mid s = 1\right) = \left(\frac{\sigma_C}{S_C}\right)\phi\left(\frac{u_C - \mu_C}{\sigma_C}\right). \tag{38}$$

This equation shows that self-selection, i.e., the extent to which college average ability decreases with college expansion, increases with σ_C, the variance of sector-neutral ability. To generate an increase in college graduates' earning dispersion of a similar magnitude as the 1960-2017 data,

ΔV_R and ΔV_H, the model infers that the variance of the ability to return to college is high, which in turn, indicates a non-negligible strength of self-selection.[29] To put a number on it, our estimated model suggests that in 1960, a 1 percentage point increase in the share of college graduates is locally associated with about a 0.24 percent decrease in the average sector-neutral ability of college graduates (specifically, $\sigma_C \phi[(u_C - \mu_C)/\sigma_C] = 0.24$).

A similar logic applies to sectoral choices. The main difference is that we do not restrict abilities to be positively selected into the two sectors. For both sectors $j \in \{H, R\}$, the direction and magnitude of self-selection are governed by the joint distribution of sector-specific abilities. To see this, express:

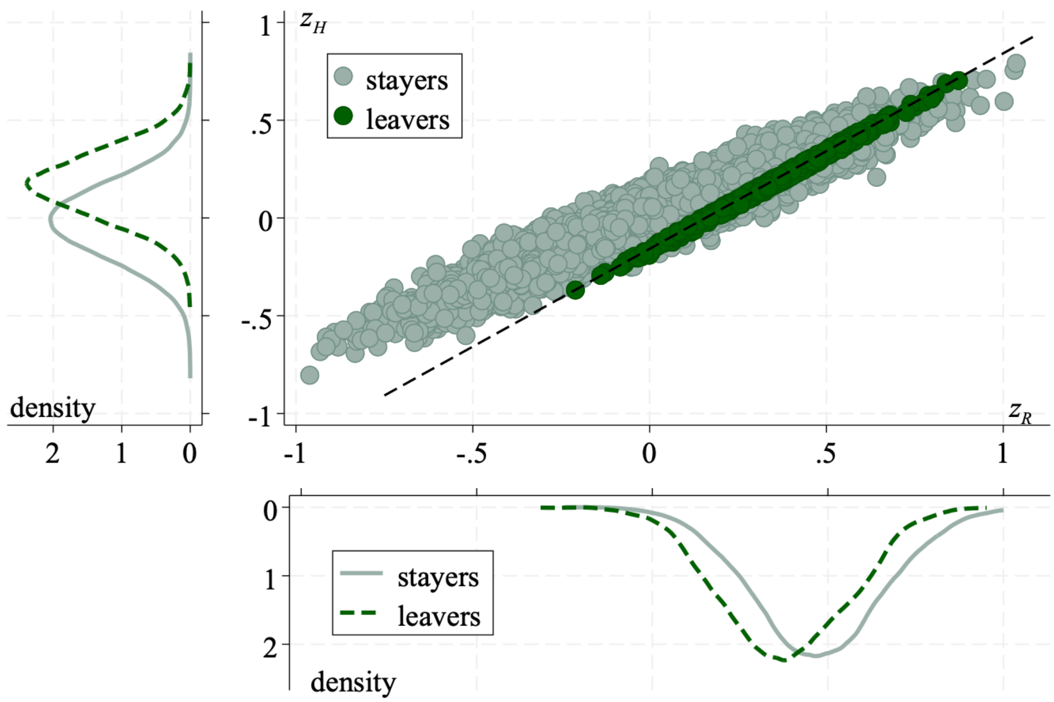
$$\mathbb{E}\left(z_{ji} \mid s = 1, j = j\right) = \left(\frac{\sigma_j^2 - \rho\sigma_j\sigma_k}{\theta\, S_{j|C}}\right)\phi\left(\frac{\tilde{w}_j - \tilde{w}_k}{\theta}\right) \tag{39}$$

The longitudinal moments are key to identifying the distribution of sector-specific abilities, (z_R, z_H), as demonstrated in Figure 13. We first plot the sector-specific abilities of college graduates simulated from our estimated model. The dashed line depicts the indifference curve, $\ln z_R + \tilde{w}_R = \ln z_H + \tilde{w}_H$; individuals on the right choose the research sector R, and individuals on the left choose the production sector H. Each light gray dot denotes 100 sector stayers defined in steps 8 and 9 of the indirect inference specified in Section 4.3, whereas each dark green dot denotes 100 leavers. By design of the indirect inference, sector leavers are the ones on the margin of the indifference curve. On the top of this figure, we plot the marginal densities of $\ln z_R$ for the stayers and leavers of the research sector respectively. It replicates the pattern we document from the 2013-2015 NSCG in Panel (a) of Figure 9: on average, leavers earn about 10% less than stayers in the research sector. Likewise, we plot the marginal density of $\ln z_H$ for stayers and leavers of the H sector on the right of the figure. It also replicates the pattern we document in Panel (b) of Figure 9: on average, leavers earn about 20% more than stayers.

The figure shows that the panel moments can only be reconciled with a positive correlation between sectoral-specific abilities, $\rho > 0$, and a larger variance of research ability, $\sigma_R^2 > \sigma_H^2$. For instance, a negative correlation would have predicted stayers to reveal higher average earnings than movers in both sectors and hence been unable to replicate the moments in the non-research sector. Given a positive correlation, if σ_H^2 became larger than σ_R^2, the relation between stayers'

[29] Though we target changes in dispersion, the resulting (initial) level of dispersion in the estimated model is 0.345 in the researcher sector and 0.348 in the non-researcher college graduate sector, compared to 0.204 and 0.398 in the data. Note that the data sources we use likely underestimate the level of dispersion due to top coding and under-sampling of the highest earners; see e.g. Figure A.2 in Song, Price, Guvenen, Bloom, and Von Wachter (2019).

Figure 13: Estimated distribution of sector-specific abilities, stayers vs. movers



and leavers' earnings would have flipped in both sectors. The panel moments, therefore, help identify the latent ability distribution.[30]

The estimated model indeed points to a relatively small effect of self-selection through sectoral choices. To put a number on it, our estimation suggests that in 1960, a 1 percentage point expansion in $S_{R|C}$ is locally associated with about a 0.02 percent decrease in the average research-specific ability. On the other hand, the abilities are negatively selected into the H sector: a 1 percentage point expansion in $S_{H|C}$ is locally associated with a 0.01 percent decrease in sector-specific ability for non-research college graduates (specifically, $(\sigma_R^2 - \rho\sigma_R\sigma_H)\phi[(\tilde{w}_R - \tilde{w}_H)/\theta]/\theta = 0.01$).

We highlight that, by targeting ΔV_R and ΔV_H, we implicitly assume that the entire change in earning dispersion comes from ability self-selection. A plausible alternative would be treating the ΔV_N in data as capturing changes in the dispersion of external earning factors that are common to all sectors and, hence, targeting the relative earning dispersion. However, such com-

[30] Sectoral shares and the first two moments of earnings are not enough to identify the distribution of latent abilities with flexible correlation (ρ) even within the class of log normal distributions. Note that we could identify ρ using a higher-order moment of the earnings distribution (e.g. skewness—see French and Taber (2011) for details); however, there is little economic intuition for why a higher-order moment identifies ρ. More practically, it is unlikely that these higher-order moments are well-measured in the data.

mon factors do not seem to be a significant component of earning dispersion in the no-college sector because V_R is much smaller than V_N. This implies that V_N and its change mostly reflects sector-specific factors, in which case, targeting the relative dispersion, e.g., $V_R - V_N$, could lead to a severe downward bias in our estimated change of ability. We hence keep the data-observed changes in earning dispersion, ΔV_R and ΔV_H, as targeted moments and acknowledge the underlying assumption. We also want to highlight that it is unclear whether the (true) change of earning dispersion due to self-selection should lie above or below the data value because the change of dispersion in other factors need not be positive.

4.5 Sensitivity of estimated parameters with respect to moments

To further investigate the mapping from data moments to model parameters imposed by our model, we implement the sensitivity measure defined in Andrews, Gentzkow, and Shapiro (2017). Sensitivity Λ is defined as:

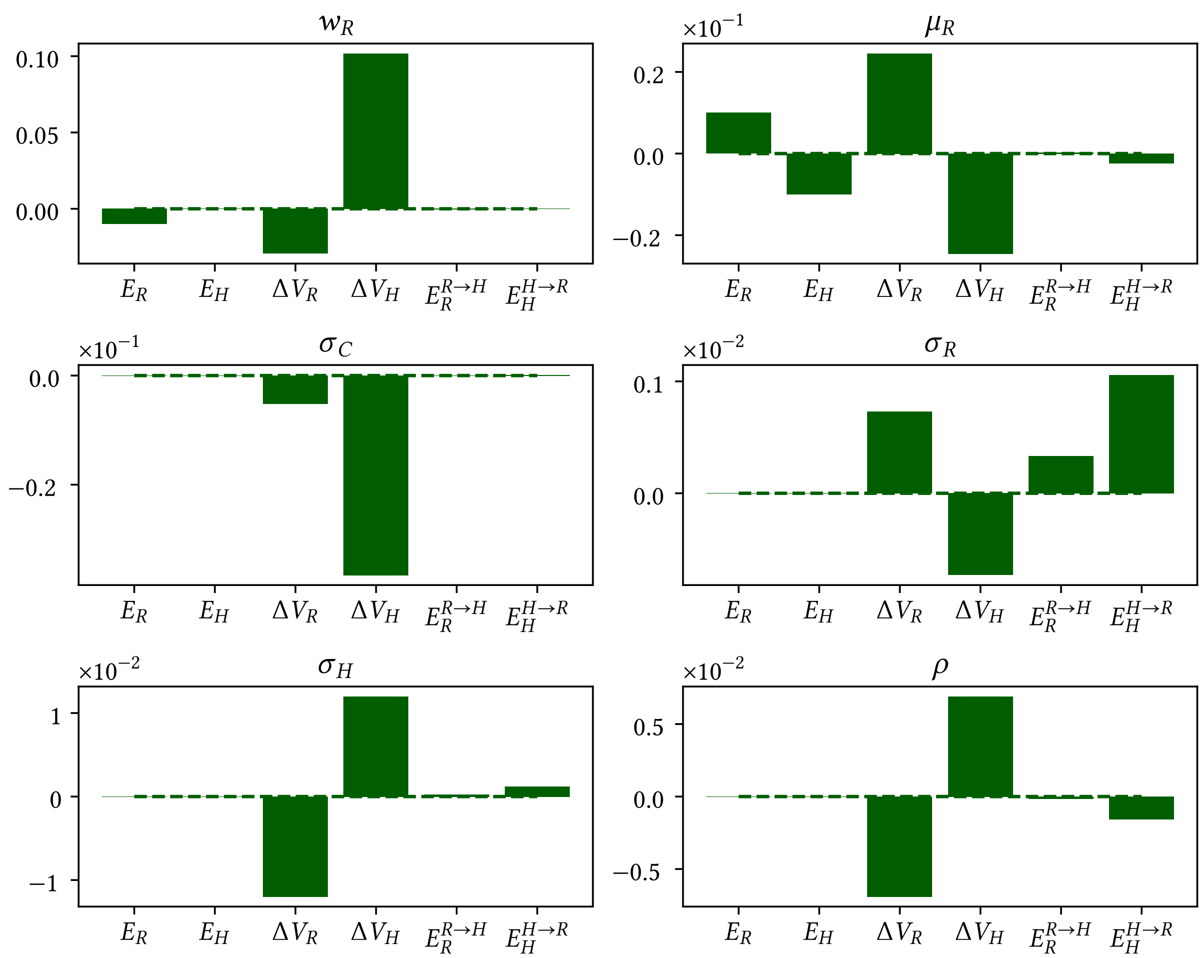
$$\Lambda = \left[G(\hat{\boldsymbol{p}})'\boldsymbol{W}G(\hat{\boldsymbol{p}})\right]^{-1} G(\hat{\boldsymbol{p}})'\boldsymbol{W}, \tag{40}$$

with $G(\hat{\boldsymbol{p}})$ the Jacobian of the vector of targeted model moments evaluated at the parameter estimate $\hat{\boldsymbol{p}}$ and $\boldsymbol{W}$ the weight matrix used in the estimation. It is termed sensitivity because it captures the first-order asymptotic bias in parameter estimates with respect to a local perturbation in the assumptions on the data-generating process. Intuitively, an element (i, j) of Λ corresponds to the sensitivity of parameter i with respect to a change in moment j. Figure 14 shows the elements of Λ under our estimated parameter vector. Each graph plots the sensitivity of a given parameter to the targeted moments. The elements of Λ are scaled so that the numbers are interpretable as the asymptotic bias in a parameter given a .01 log point change in a particular moment.[31] Note that some sensitivities are sufficiently close to 0 that they effectively disappear.

In line with our discussion in Section 4.4, the standard deviations of ability $(\sigma_C, \sigma_R, \sigma_H)$ are particularly sensitive to the change in the variance of log earnings among researchers and among non-researcher college graduates. The standard deviation of the common component of ability across the two college sectors (σ_C) is particularly sensitive to the change in the variance of ability among college graduates, presumably because most college graduates are not researchers. The sensitivity of parameters to the panel moments incorporated in our estimation is highest

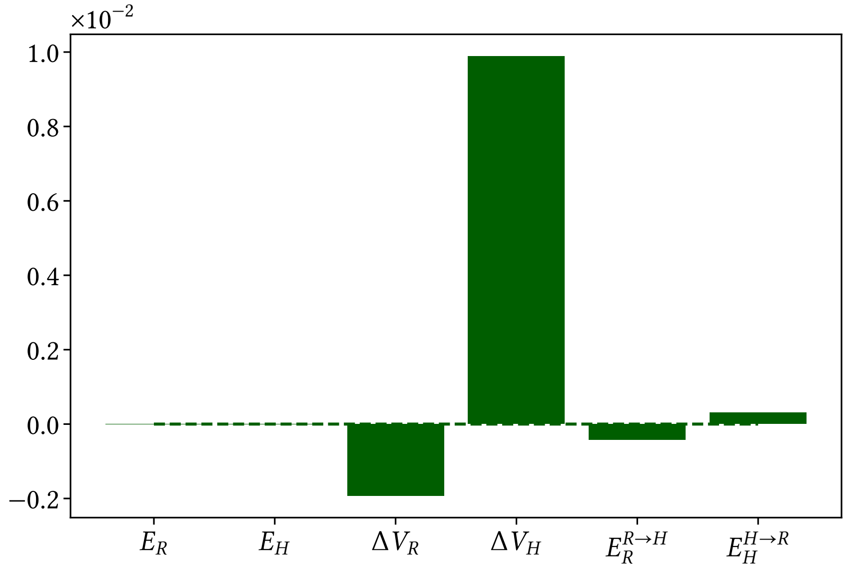
[31] Because the objective function used to estimate the model belongs to the class of classical minimum distance estimators as defined in Andrews, Gentzkow, and Shapiro (2017), the most relevant set of perturbations to consider are additive shifts to the moment functions, in which case the asymptotic bias under a given set of alternative assumptions resulting in additive shifts to the moment functions η is simply $\eta \times \Lambda$.

Figure 14: Sensitivity of parameters to numerically-approached moments



Notes: Each bar plots an element of the sensitivity matrix Λ as defined in equation (40). The height of the bar measures the sensitivity of an estimated model parameter to a moment targeted in the estimation. All elements are divided by 100 so that they capture the sensitivity of a parameter with respect to a 0.01 log point change in a given moment.

Figure 15: Sensitivity of estimated change in ability among researchers to targeted moments



Notes: Each bar plots the sensitivity of the estimated change of ability among researchers to a targeted moment. All elements are divided by 100 so that they capture the sensitivity of a parameter with respect to a 0.01 log point change in a given moment.

for the distributional parameters of the noncommon portion of ability among college graduates $(\sigma_R, \sigma_H, \rho)$; this is again in line with our discussion in 4.4. Finally, note that none of the parameters are particularly sensitive to average log earnings among researchers E_R and non-researcher college graduates E_H.

To summarize the estimated sensitivities' influence on the model outcome we focus on in this paper, we can further calculate the sensitivity of the estimated change in ability among researchers, which by the delta method is just $C(\hat{\boldsymbol{p}}) \times \Lambda$ with $C(\hat{\boldsymbol{p}})$ the Jacobian of the change in ability among researchers with respect to the parameter vector, evaluated at the estimated parameters. Figure 15 plots the resulting vector. The estimated change in ability among researchers is particularly sensitive to the change in the variance of log earnings among non-researcher college graduates; this follows because the model infers almost the entire decline in ability as coming from the education margin of the model, the strength of which is controlled by σ_C. We want to emphasize that this is only true locally: if the data moments were such that the model attributed a greater role to the sectoral choice channel, then the sensitivities on the panel moments and the change in the variance of log earnings among researchers would be magnified. The figure should thus not be interpreted as saying that moments other than the change in the variance of log earnings among non-researcher college graduates are irrelevant in pinning down the change in ability among researchers; indeed even locally these sensitivities are non-negligible.

4.6 External validation for self-selection in college education

Our model features weak self-selection into research among college graduates, but comparatively strong self-selection into education. Following the discussion in Section 4.5, the model infers this strong selection into education primarily by the large change in the variance of log earnings in the college-educated sectors, together with the longitudinal patterns discussed in Section 3.3. An underlying assumption of our estimation is that the entire change in the variance of log earnings in the college-educated sectors stems from self-selection. Large deviations from this assumption could of course lead to bias in our estimated parameters.

To gauge how concerned we should be about this, we can compare the strength of selection into education implied by our model to existing estimates in the literature. Carneiro and Lee (2011) argue that the college wage premium in 2010 would have been around 30% higher if the average ability of college graduates had not declined between 1960 and 2000; our model predicts that the college wage premium would have been around 32% higher if the average ability of college graduates had not declined over the same time period. [32]

Another approach to comparing the strength of self-selection in our model to estimates from the literature is to compare the underlying distributions. Because ability has no natural units, and because the exact interpretation of ability differs across models, we do this by comparing the difference in average ability between those with and without a college degree scaled by the standard deviation of the ability distribution. Our model predicts that the average z_{Ci} of college graduates is around 1.86 standard deviations higher than the average z_{Ci} of non-college-graduates in 2017, and around 1.93 standard deviations higher in 1960.[33] Hendricks and Schoellman (2014) estimate the difference in ability between college graduates and high-school graduates without a college degree to be between 1.44 and 3.75 standard deviations depending on specification. Similarly, Heckman, Humphries, and Veramendi (2018) estimate the difference in average cognitive ability between college graduates and high-school graduates without a college degree to be about 2

[32]Other studies measure the strength of self-selection by the proportion ofthe college wage premium (rather than the change of premium) that can be attributed to differences in ability e.g. Hendricks and Schoellman (2014), Hendricks and Leukhina (2017). Comparing these estimates to those produced by our model is difficult because we do not model heterogeneity in the non-college educated sector, and are thus likely to strongly overstate the returns to schooling.

[33]Note that, because of independence, there is no difference of mean between those with and without a college degree in the sector-specific abilities.

standard deviations. If anything, our estimate of the strength of selection into education is thus on the conservative side.

5 Revisiting the permanent diminishing returns of research

Given the estimated model, we can compute the log average ability in the research sector between 1960 and 2017. Our model suggests that, under our estimated ability distribution, workers with a comparative advantage in the research sector also tend to have an absolute advantage. Therefore, workers newly entering the expanding research sector have lower research ability compared with researchers in 1960, and hence the average research ability falls as the researcher share grows. Over the period 1960-2017, our model suggests that average researcher ability, $\overline{Z}_R$, has decreased by about 48.9%, implying an annual change rate of about -0.86%. Quantitatively, this composition effect is mainly from self-selection in education. The self-selection in sectoral choices, on the contrary, slightly drives up the average researcher's ability because abilities are positively self-selected into research, and the share of researchers among college graduates declines.

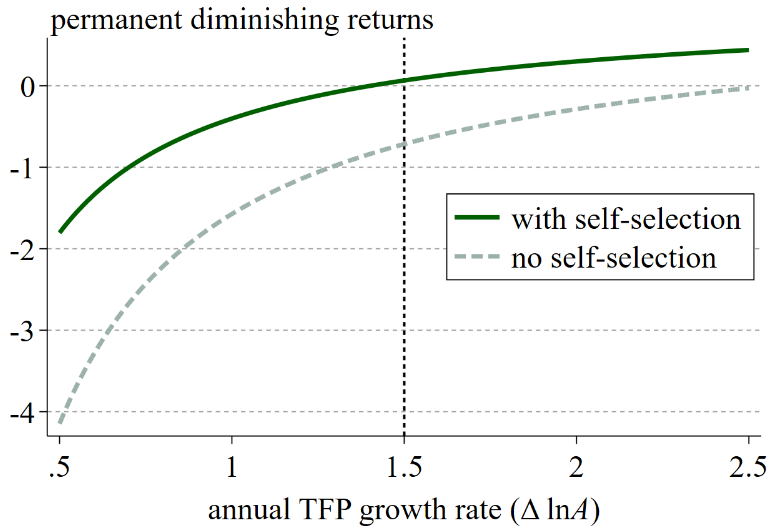
We can gauge the importance of this fall by returning to the decomposition in Section 2. Recall that we can write ϕ as:

$$\phi = 1 - \frac{\Delta \ln L_R + \Delta \ln \overline{Z}_R}{\Delta \ln A}. \tag{41}$$

Given that the growth rate of the number of researchers is about 2.6% per year, Figure 16 plots the implied values of ϕ for a range of plausible annual growth rates of TFP. The light dashed line shows the value of ϕ without considering self-selection, i.e., holding $\Delta \ln \overline{Z}_R = 0$; whereas the solid line allows $\overline{Z}_R$ to fall as inferred by our estimated model. We see that for any value for the average TFP growth rate, the value of ϕ increases significantly when taking self-selection into account. This also implies that the estimated fall in $\overline{Z}_R$ constitutes a significant portion of the overall fall in R&D productivity.

Take a TFP growth rate of 1.5% as a representative example. Accounting for self-selection moves the value of ϕ from around -0.72 to around 0.07. Recall that the long-run growth predictions are $g_y \propto g_n/(1-\phi)$ in general semi-endogenous growth models and $g_y = g_n/(1-\phi)$ in Jones (1995). This means that, through the lens of these models, the predicted long-run growth rate becomes about 1.85 times larger when taking self-selection into account. Furthermore, if the population grows at 1% per year, the long-run growth rate of income implied by Jones (1995)

Figure 16: Model implied permanent diminishing return (ϕ)



Notes: Figure reports the permanent diminishing returns of research (ϕ) implied by different annual rates of TFP growth($\Delta \ln A$). The solid line depicts the ϕ calculated by (41), whereas the dashed line depicts the ϕ calculated without considering changes in researcher ability, i.e. assuming $\Delta \ln \overline{Z}_R = 0$.

would increase from around 0.58% to 1.08%. In turn, this implies that, on the balanced growth path, the ratio in per capita incomes between these two hypothetical economies would approximately double every 138 years.

Recall that the scale elasticity is also a key determinant of the speed of convergence to the balanced growth path; fixing research intensity, the half life of convergence to the balanced growth rate given some initial rate is again proportional to $1/1 - \phi$ (Jones, 2022).

6 Conclusion

The permanent returns of research, as captured by the scale elasticity of existing knowledge in producing the flow of new knowledge, is a key component in determining long-run growth rates in R&D-based growth theories. Previous studies have estimated its value by attributing all observed declines in researcher productivity to permanent diminishing returns. In this paper, we investigate transitional source of diminishing returns of research, namely a drop in the average quality of researchers caused by the expansion of the research sector and self-selection in researcher abilities. We argue that ignoring this source of transitional diminishing returns biases the estimated scale of permanent diminishing returns downwards, and hence understates the predicted long-run growth rate of the economy.

Our approach to measuring the transitional diminishing returns of research consists of two steps. First, we measure the time trend of researchers' workforce and their earnings distribution in the United States over the period 1960-2017. Our measurements leverage unique information about the work activities of college-educated labor provided by the NSCG. We then impute the share of researchers at a detailed occupational level to the Decennial Census and the ACS to construct aggregate moments. Our result confirm the fact that the overall share of researchers has been growing over time. In addition, we find that the share of researchers among college-educated labor has been declining and that the dispersion of earnings for researchers and college-degree workers has been rising over time. These facts are aligned with the prediction of models where workers self-select into education and sectors on the basis of a ability, pointing to a potential decline in the average quality of researchers—the transitional diminishing returns of research.

Our second step is to estimate a model of researcher supply, where individuals are heterogeneous along three dimensions— their expected return on a college education, their ability in the research sector, and their ability in the non-researcher college-graduate sector. In this framework, the key determinant for the direction and magnitude of self-selection is the joint distribution of these individual-specific types, which captures the relation between comparative advantage and absolute advantage. The purpose of formalizing self-selection in a model is to quantify the transitional diminishing returns in conjunction with our measured trends. We thus conduct a quantitative analysis where the distribution of abilities is parameterized using a joint log-normal distribution. Given the estimated parameters, our baseline model infers a 0.86% annual decline rate in the average ability of researchers from 1960-2017 in the United States. Separating the transitional diminishing returns moves the estimated permanent diminishing returns of R&D from -0.72 to about 0.07. Through the lens of semi-endogenous growth models, this change predicts a 1.85 higher growth rate in the long-run.

References

Akcigit, Ufuk, Jeremy Pearce, and Marta Prato (2020). "Tapping into Talent: Coupling Education and Innovation Policies for Economic Growth." NBER Working Paper 27862 .

Andrews, Isaiah, Matthew Gentzkow, and Jesse M Shapiro (2017). "Measuring the sensitivity of parameter estimates to estimation moments." *The Quarterly Journal of Economics* 132 (4):1553–1592.

Atkeson, Andrew and Ariel Burstein (2019). "Aggregate Implications of Innovation Policy." *Journal of Political Economy* 127 (6):2625–2683.

Barro, Robert J. and Jong Wha Lee (2013). "A New Data Set of Educational Attainment in the World, 1950–2010." *Journal of Development Economics* .

Bloom, Nicholas, Charles I. Jones, John Van Reenen, and Michael Webb (2020). "Are Ideas Getting Harder to Find?" *American Economic Review* 110 (4):1104–1144.

Card, David (1999). "Chapter 30 - The Causal Effect of Education on Earnings." In *Handbook of Labor Economics*, vol. 3, edited by Orley C. Ashenfelter and David Card. Elsevier, 1801–1863.

Carneiro, Pedro and Sokbae Lee (2011). "Trends in Quality-Adjusted Skill Premia in the United States, 1960-2000." *American Economic Review* 101 (6):2309–49.

Feenstra, Robert C., Robert Inklaar, and Marcel P. Timmer (2015). "The Next Generation of the Penn World Table." *American Economic Review* 105 (10):3150–82.

Fernald, John G. and Charles I. Jones (2014). "The Future of US Economic Growth." *American Economic Review: Papers & Proceedings* 104 (5):44–49.

French, Eric and Christopher Taber (2011). "Identification of Models of the Labor Market." In *Handbook of Labor Economics*, vol. 4a, edited by Orley Ashenfelter and David Card. Amsterdam: Elsevier, 537–617.

Heckman, James J. and Bo E. Honoré (1990). "The Empirical Content of the Roy Model." *Econometrica* 58 (5):1121–1149.

Heckman, James J., John Eric Humphries, and Gregory Veramendi (2018). "Returns to Education: The Causal Effects of Education on Earnings, Health, and Smoking." *Journal of Political Economy* 126 (S1):S197–S246. URL https://doi.org/10.1086/698760.

Heckman, James J. and Guilherme L. Sedlacek (1985). "Heterogeneity, Aggregation, and Market Wage Functions: An Empirical Model of Self-Selection in the Labor Market." *Journal of Political Economy* 93 (6):1077–1125.

Heckman, James J. and Guilherme L. Sedlacek (1990). "Self-Selection and the Distribution of Hourly Wages." *Journal of Labor Economics* 8 (1):S329–S363.

Hendricks, Lutz and Oksana Leukhina (2017). "How risky is college investment?" *Review of Economic Dynamics* 26:140–163.

Hendricks, Lutz and Todd Schoellman (2014). "Student abilities during the expansion of US education." *Journal of Monetary Economics* 63:19–36.

Hsieh, Chang-Tai, Erik Hurst, Charles I. Jones, and Peter J. Klenow (2019). "The Allocation of Talent and U.S. Economic Growth." *Econometrica* 87 (5):1439–1474.

Jones, Charles I. (1995). "R&D-Based Models of Economic Growth." *Journal of Political Economy* 103 (4):759–784.

Jones, Charles I. (2002). "Sources of U.S. Economic Growth in a World of Ideas." *American Economic Review* 92 (1):220–239.

Jones, Charles I. (2005). "Growth and Ideas." In *Handbook of Economic Growth*, vol. 1B, edited by Philippe Aghion and Steven N. Durlauf. Amsterdam: Elsevier, 1063–1111.

Jones, Charles I. (2016). "The Facts of Economics Growth." In *Handbook of Macroeconomics*, vol. 2A, edited by John B. Taylor and Harald Uhilg. North Holland, 3–69.

Jones, Charles I (2022). "The past and future of economic growth: A semi-endogenous perspective." *Annual Review of Economics* 14 (1):125–152.

Lagakos, David and Michael E. Waugh (2013). "Selection, Agriculture, and Cross-Country Productivity Differences." *American Economic Review* 103 (2):948–980.

Peters, Michael (2019). "Heterogeneous Markups, Growth and Endogenous Misallocation." *Working Paper* .

Romer, Paul (1990). "Endogenous Technological Change." *Journal of Political Economy* 98 (5):S71–S102.

Roy, A. D. (1951). "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers New Series* 3:135–146.

Song, Jae, David J Price, Fatih Guvenen, Nicholas Bloom, and Till Von Wachter (2019). "Firming up inequality." *The Quarterly journal of economics* 134 (1):1–50.

Willis, Robert J. and Sherwin Rosen (1979). "Education and Self-Selection." *Journal of Political Economy* 87 (5):S7–S36.

Wolfe, Raymond M. (2014). "Business R&D performance in the United States tops $300 billion in 2012." NCSES Info Brief 15303, National Science Foundation.

Young, Alwyn (2014). "Structural Transformation, the Mismeasurement of Productivity Growth, and the Cost Disease of Services." *American Economic Review* 104 (11):3635–3667.

Appendix

A Full semi-endogenous growth model

This section describes the demand-side of the semi-endogenous growth model described in Section 2, and relates it to a semi-endogenous growth model without selection into research. To allow for decentralization of the model, first suppose that the stock of ideas A_t allows for the production of a continuum of intermediate input varieties with measure equal to A_t. Entry into research is free and discovering an idea enables an entrant to obtain a perpetual patent to the intermediate input variety corresponding to that idea. In the final goods sector, a representative firm uses low-skilled production labor, high-skilled production labor, and the continuum of intermediate input varieties to produce the final consumption good; its profit-maximization problem is thus

$$\max_{x_{it}, Z_{H,t}, Z_{N,t}} \frac{1}{1-\alpha-\beta} \int_0^{A_t} x_{i,t}^{1-\alpha-\beta} di Z_{H,t}^{\beta} Z_{N,t}^{\alpha} di - W_{H,t} Z_{H,t} - W_{N,t} Z_{N,t} - \int_0^{A_t} p_{i,t} x_{i,t} di. \quad \text{(A1)}$$

The aggregate production function thus takes a simple Cobb-Douglas form; this assumption is inessential but simplifies the characterization of the model significantly. Conditional on entry, the intermediate input monopolists can produce their variety at constant marginal cost ψ. The consumer-side is standard: a representative household with discount rate ρ and instantaneous utility function $\frac{C_t^{1-\sigma}-1}{1-\sigma}$ chooses consumption and savings given their budget constraint and real interest rate r_t. Let $\mathcal{R}_t = \frac{W_{R,t}}{W_{H,t}}, \mathcal{R}_t^N = \frac{W_{R,t}}{W_{N,t}}$, then we can summarize the model with following set of equations:

$$\text{Idea production} \quad \dot{A}_t = \chi A_t^{\phi} Z_{R,t} \quad \text{(A2)}$$

$$\text{Researcher labor supply} \quad Z_{R,t}^S = L_t \Gamma_R \left(\mathcal{R}_t, \mathcal{R}_t^N\right) \quad \text{(A3)}$$

$$\text{High-skill production labor supply} \quad Z_{H,t}^S = L_t \Gamma_H \left(\mathcal{R}_t, \mathcal{R}_t^N\right) \quad \text{(A4)}$$

$$\text{Low-skill production labor supply} \quad Z_{N,t}^S = L_t \Gamma_N \left(\mathcal{R}_t, \mathcal{R}_t^N\right) \quad \text{(A5)}$$

$$\text{Optimal intermediate input price} \quad p_{i,t} = \frac{\psi}{1-\beta} \quad \text{(A6)}$$

$$\text{Euler equation} \quad \frac{\dot{C}_t/\dot{L}_t}{C_t/L_t} = \frac{1}{\sigma}(r_t - \rho) \quad \text{(A7)}$$

$$\text{High-skill production labor demand} \quad Z_{H,t}^{D} = \left[\frac{\beta Z_{N,t}^{\alpha}\tilde{A}_t}{1-\alpha-\beta}\right]^{\frac{1}{1-\beta}} W_{H,t}^{-\frac{1}{1-\beta}} \tag{A8}$$

$$\text{Low-skill production labor demand} \quad Z_{N,t}^{D} = \left[\frac{\alpha Z_{H,t}^{\beta}\tilde{A}_t}{1-\alpha-\beta}\right]^{\frac{1}{1-\alpha}} W_{N,t}^{-\frac{1}{1-\alpha}} \tag{A9}$$

$$\text{Intermediate input demand} \quad x_{i,t}^{D} = Z_{N,t}^{\frac{\alpha}{\alpha+\beta}} Z_{H,t}^{\frac{\beta}{\alpha+\beta}} p_{i,t}^{-\frac{1}{\alpha+\beta}} \tag{A10}$$

$$\text{Free entry into research} \quad W_{R,t} = \chi A_t^{\phi} V_{i,t} \tag{A11}$$

with $V_{i,t}$ the value of the perpetual patent for intermediate input i at time t, $\tilde{A}_t \equiv \int_0^{A_t} x_{i,t}^{1-\alpha-\beta} di$, and $\Gamma_j(\cdot)$ denoting the function that maps the joint distribution of (z_{Ci}, z_{Ri}, z_{Hi}) to labor productivity in sector j. An equilibrium is then a set of allocations $\{Z_{N,t}, Z_{R,t}, Z_{N,t}, C_t\}$, a time path of ideas A_t, and a set of prices $\{p_{i,t}, r_t, W_{R,t}, W_{H,t}, W_{N,t}\}$ that satisfy these equations.

A.1 Solving the balanced growth path

Normalizing $p_{i,t} = 1$, the instantaneous profit function is

$$\pi_t = Z_{H,t}^{\frac{\beta}{\alpha+\beta}} Z_{N,t}^{\frac{\alpha}{\alpha+\beta}} (1-\psi). \tag{A12}$$

We can also solve

$$\tilde{A}_t = Z_{N,t}^{\frac{\alpha(1-\alpha-\beta)}{\alpha+\beta}} Z_{H,t}^{\frac{\beta(1-\alpha-\beta)}{\alpha+\beta}} A_t \tag{A13}$$

Define a balanced growth path (BGP) to be a path of allocations under which consumption and output grow at the same constant rate, and denote with $*$ the values of equilibrium objects along this BGP. Suppose, for now, that $\frac{\mu_{R,t}}{\mu_{H,t}} = \frac{\mu_{N,t}}{\mu_{R,t}\mu_{C,t}} = 1$; once we have solved the BGP under this assumption, relaxing it is straightforward. On a BGP, the Hamilton-Jacobi-Bellman equation must satisfy

$$(r^* - g_L)V_{i,t}^* = \pi_t^*. \tag{A14}$$

Combining these equations with the labor demand equations stemming from the final goods producer's profit maximization problem yields:

$$W_{R,t}^* = \frac{\chi(1-\psi)}{r^* - g_L} A_t^{*\phi} Z_{H,t}^{*\frac{\beta}{\alpha+\beta}} Z_{N,t}^{*\frac{\alpha}{\alpha+\beta}} \tag{A15}$$

$$W_{H,t}^* = \frac{\beta}{1-\alpha-\beta} Z_{N,t}^{*\frac{\alpha}{\alpha+\beta}} Z_{H,t}^{*-\frac{\alpha}{\alpha+\beta}} A_t^* \tag{A16}$$

$$W_{N,t}^* = \frac{\alpha}{1-\alpha-\beta} Z_{N,t}^{*-\frac{\beta}{\alpha+\beta}} Z_{H,t}^{*\frac{\beta}{\alpha+\beta}}. \tag{A17}$$

Along a balanced growth path, the employment shares of the different types of labor must be constant; we therefore have

$$\mathcal{R}^* = \frac{W_{R,t}^*}{W_{H,t}^*} = \frac{\chi(1-\alpha-\beta)(1-\psi)}{\beta(r^* - g_L)} A_t^{*\phi-1} Z_{H,t}^* \tag{A18}$$

$$\mathcal{R}_N^* = \frac{W_{R,t}^*}{W_{N,t}^*} = \frac{\chi(1-\alpha-\beta)(1-\psi)}{\alpha(r^* - g_L)} A_t^{*\phi-1} Z_{N,t}^* \tag{A19}$$

Rearranging the idea production function yields $A_t^{*\phi-1} = g_A^* \left[\chi Z_{R,t}^*\right]^{-1}$ and thus we can rewrite the above as:

$$\mathcal{R}^* = \frac{(1-\alpha-\beta)(1-\psi)}{\beta(r^* - g_L)} \frac{Z_{H,t}^*}{Z_{R,t}^*} \tag{A20}$$

$$\mathcal{R}_N^* = \frac{(1-\alpha-\beta)(1-\psi)}{\alpha(r^* - g_L)} \frac{Z_{N,t}^*}{Z_{R,t}^*}. \tag{A21}$$

Finally, plugging in the labor supply equations:

$$\mathcal{R}^* = \frac{(1-\alpha-\beta)(1-\psi)}{\beta(r^* - g_L)} \frac{\Gamma_H(\mathcal{R}^*, \mathcal{R}_N^*)}{\Gamma_R(\mathcal{R}^*, \mathcal{R}_N^*)} \tag{A22}$$

$$\mathcal{R}_N^* = \frac{(1-\alpha-\beta)(1-\psi)}{\alpha(r^* - g_L)} \frac{\Gamma_N(\mathcal{R}^*, \mathcal{R}_N^*)}{\Gamma_R(\mathcal{R}^*, \mathcal{R}_N^*)}. \tag{A23}$$

To solve for r^*, note that we can rewrite the idea production function as

$$\frac{\dot{A}_t^*}{A_t^*} = \chi A_t^{*\phi-1} Z_{R,t}^* \tag{A24}$$

Taking logs, deriving with respect to time, and rearranging yields

$$g_A^* = \frac{g_{z_{Ri}}^*}{1-\phi} = \frac{g_L}{1-\phi} \tag{A25}$$

Plugging this into the Euler equation gives

$$r^* = \frac{\sigma}{1-\phi} g_L + \rho, \tag{A26}$$

and so equations (A22) and (A23) can be rewritten

$$\mathcal{R}^* = \frac{(1-\alpha-\beta)(1-\psi)}{\beta(\frac{\sigma-(1-\phi)}{\phi} g_L + \rho)} \frac{\Gamma_H(\mathcal{R}^*, \mathcal{R}_N^*)}{\Gamma_R(\mathcal{R}^*, \mathcal{R}_N^*)} \tag{A27}$$

$$\mathcal{R}_N^* = \frac{(1-\alpha-\beta)(1-\psi)}{\alpha(\frac{\sigma-(1-\phi)}{\phi} g_L + \rho)} \frac{\Gamma_N(\mathcal{R}^*, \mathcal{R}_N^*)}{\Gamma_R(\mathcal{R}^*, \mathcal{R}_N^*)}. \tag{A28}$$

This yields two equations in two unknowns, which can be solved for $(\mathcal{R}^*, \mathcal{R}_C^*)$. Given this solution, the BGP allocations of all equilibrium objects can be found by unwinding the set of equation substitutions leading to equations (A27) and (A28). Note that this model has the exact same BGP growth rate as one without selection into research, thus justifying our description of selection effects as transitional. This can easily be verified by setting $z_{Ri} = z_P = z_{Ci} = 1$ for all workers and $\mu_{C,t} = 0$, in which case there is no self-selection but equation (40) still holds.

To relax the assumption that $\frac{\mu_{R,t}}{\mu_{H,t}} = \frac{\mu_{N,t}}{\mu_{R,t}\mu_{C,t}} = 1$, suppose that $\frac{\mu_{R,t}}{\mu_{H,t}} = a_t$, $\frac{\mu_{N,t}}{\mu_{R,t}\mu_{C,t}} = b_t$ and let $**$ denote values of equilibrium objects in the BGP which holds under this assumption. Then if

$$\frac{W_{R,t}^{**}}{W_{H,t}^{**}} = \frac{W_{R,t}^{*}}{W_{H,t}^{*}} \times \frac{1}{a_t} \tag{A29}$$

$$\frac{W_{R,t}^{**}}{W_{N,t}^{**}} = \frac{W_{R,t}^{*}}{W_{N,t}^{*}} \times \frac{1}{b_t} \tag{A30}$$

we have $\mathcal{R}^{**} = \mathcal{R}^*$, $\mathcal{R}_N^{**} = \mathcal{R}_N^*$ and so, with the exception of the relative wage rates, the BGP is unchanged.[A1]

[A1] Clearly this BGP will break down if either $a_t \to \infty$ or $b_t \to \infty$.

A.2 Evolution of average log ability without imposing independence

In the main text, we derive equation (12) assuming that z_{Ci} is independent of (z_{Ri}, z_{Hi}). We drop this assumption here.

$$\begin{aligned}\bar{z}_t^R &= \mathbb{E}\left[z_{Ci} + z_{Ri} \mid z_{C,t}^m \leq z_{Ci}, z_{R,t}^m(z_H) \leq z_{Ri}\right] \\ &= \int_0^\infty f(z_{Hi}) \int_{\tilde{w}_{N,t}-z_{Hi}}^\infty f(z_{Ci}|z_{Hi}) \int_{z_{Hi}-\tilde{w}_{R,t}}^\infty z_{Ri} f(z_{Ri}|z_{Ci}, z_{Hi}) dz_{Ri} dz_{Ci} dz_{Hi} + \\ &\int_0^\infty f(z_{Hi}) \int_{z_{Hi}-\tilde{w}_{R,t}}^\infty f(z_{Ri}|z_{Hi}) \int_{\tilde{w}_{N,t}-\tilde{w}_{R,t}-z_{Ri}}^\infty z_{Ci} f(z_{Ci}|z_{Ri}, z_{Hi}) dz_{Ri} dz_{Ci} dz_{Hi}\end{aligned} \tag{A31}$$

Taking a time derivative using Leibniz's rule we get:

$$\begin{aligned}\dot{\bar{z}}_t^R = &-\dot{\tilde{w}}_{N,t}\mathbb{E}\left[z_{Ri} \mid z_{Ci} = z_{C,t}^m, z_{Ri} > z_{R,t}^m(z_H)\right] + \dot{\tilde{w}}_{R,t}\mathbb{E}\left[z_{R,t}^m(z_H) \mid z_{Ci} > z_{C,t}^m\right] + \\ &\dot{\tilde{w}}_{R,t}\mathbb{E}\left[z_{Ci} \mid z_{Ri} = z_{R,t}^m(z_H), z_{Ci} > z_{C,t}^m\right] + \left(\dot{\tilde{w}}_{R,t} - \dot{\tilde{w}}_{N,t}\right)\mathbb{E}\left[z_{C,t}^m \mid z_{Ri} > z_{R,t}^m(z_H)\right]\end{aligned} \tag{A32}$$

If z_{Ci} is independent of (z_{Ri}, z_{Hi}) we have

$$\mathbb{E}\left[z_{Ri} \mid z_{Ci} = z_{C,t}^m, z_{Ri} > z_{R,t}^m(z_H)\right] = \mathbb{E}\left[z_{Ri} \mid z_{Ri} > z_{R,t}^m(z_H)\right] \tag{A33}$$

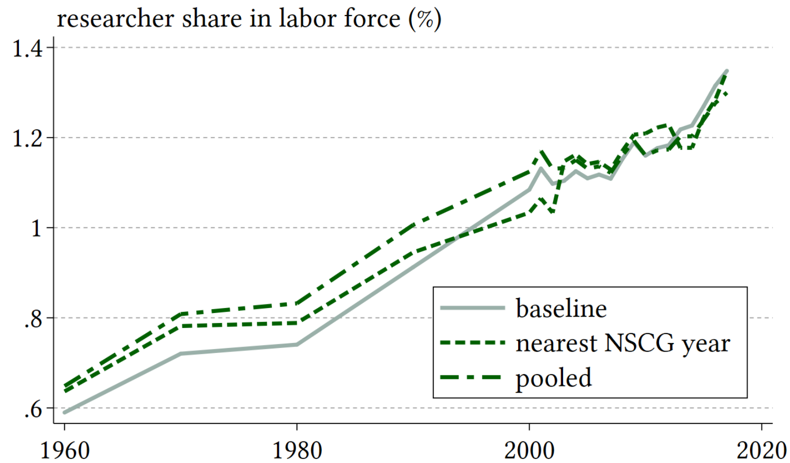
$$\mathbb{E}\left[z_{Ci} \mid z_{Ri} = z_{R,t}^m(z_H), z_{Ci} > z_{C,t}^m\right] = \mathbb{E}\left[z_{Ci} \mid z_{Ci} > z_{C,t}^m\right] \tag{A34}$$

and thus equation (12) follows.

B NSCG-Census/ACS occupation code crosswalk

A key part of our empirical approach consists of imputing occupation researcher shares from the NSCG to the Census/ACS. To do this, we have developed an occupational code crosswalk that matches each occupation in the NSCG with one or more occupational codes in the Census/ACS. The occupational codes in these data sets have a significant amount of overlap; however, the NSCG has finer codes for certain science and engineering occupations, whereas the Census/ACS has much finer codes for occupations that presumably do not employ many college graduates. An example of the former is that the NSCG has separate codes for "Physicists" and "Astronomers", whereas the Census/ACS has a single code for "Physicists and Astronomers." An example of the latter is that Census/ACS occupations such as "Police Officers," "Firefighters," and "Animal Control" are all matched to "Protective Services" in the NSCG. Note that the fact that the Census/ACS has many more occupational codes than the NSCG is irrelevant to our results

Figure A1: Different imputation procedures



Notes: Source: authors' calculation from NSCG and Census/ACS.

because we restrict the share of researchers in occupations not considered to be "science and engineering" or "science and engineering related" to be 0. The full crosswalk can be downloaded at https://drive.google.com/file/d/1-ALaOZ85Xg8rK-Z8L4XuVcs3ebFXekoZ/view?usp=sharing.

C Robustness: change of occupational researcher shares overtime

Recall that our baseline imputation procedure works as follows: first, for each occupation in the NSCG, we calculate the growth rate of the number of researchers and the number of non-researchers using all publicly-available years of the NSCG. Assuming that these growth rates are constant within each occupation, we then extrapolate backward in time using the fitted values to create a time series of the share of researchers within each occupation. Finally, we match each Census/ACS occupation with an NSCG occupation and impute our share series into the Census/ACS.

This approximation relies on the assumption that within detailed occupations, changes in the share of researchers are stable over time. Given that we only have six years of the NSCG available (covering 24 years), this might not be an ideal approximation. To test whether our results are sensitive to this assumption, we recalculate the share of researchers using two different imputation procedures. The first pools all available NSCG years and treats them as a single cross-section when computing the share of researchers within each occupation. The second imputes to each Census/ACS year the occupational share of researchers from the *closest* NSCG year, e.g., 1993 for the 1960 Census, 2003 for the 2001 ACS). Figure A1 shows the resulting time series. Though there are some level differences, the three imputation procedures result in very similar patterns.

This is because the occupational-level shares of researchers are indeed stable over time in our NSCG sample years. Figure A2 plots the occupational-level researcher shares, S_j^{NSCG}, in 1993 and 2003 against their values in 2017. Each circle represents an occupation, and the circle size shows the number of college workers in the occupation. The circles are fairly close to the 45^o line. This suggests that the cross-time change of researcher shares within occupations is negligible compared with cross-sectional variation across occupations.

D Comparison of occupational researcher shares

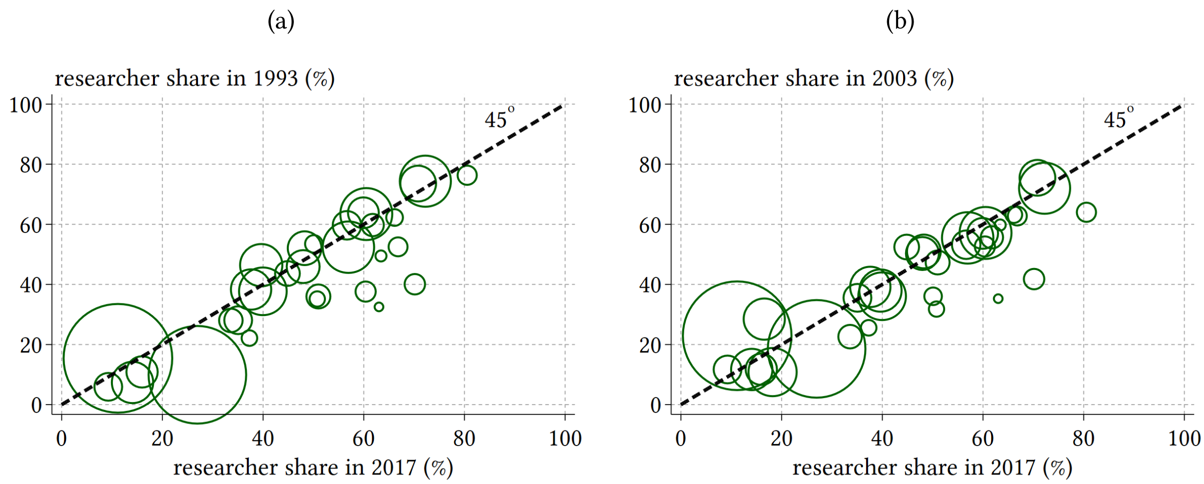
Figure A3 contrasts the occupational share of log earnings by researchers, in panel (a), and that of log earning squares, in panel (b), against the researcher share of the workforce. The dashed line depicts the 45-degree reference line. The figure indicates that the three occupational shares of researchers are nearly identical. This tells us that within detailed occupations, researchers' earnings are not systematically different from other college graduates', suggesting that within-occupation differences between researchers and non-researchers are likely not a significant source of error in our construction of aggregate trends.

E Trends of sectoral mean and variances of log earnings

In Section 3.2, we report the variance of log earnings of the two college-requiring sectors. Figure A4, in addition, reports the variance of log earnings for workers without college degrees: V_N. We see that the earnings dispersion for non-college workers in the U.S. also increases over time, though its magnitude is smaller than that of the other two sectors.

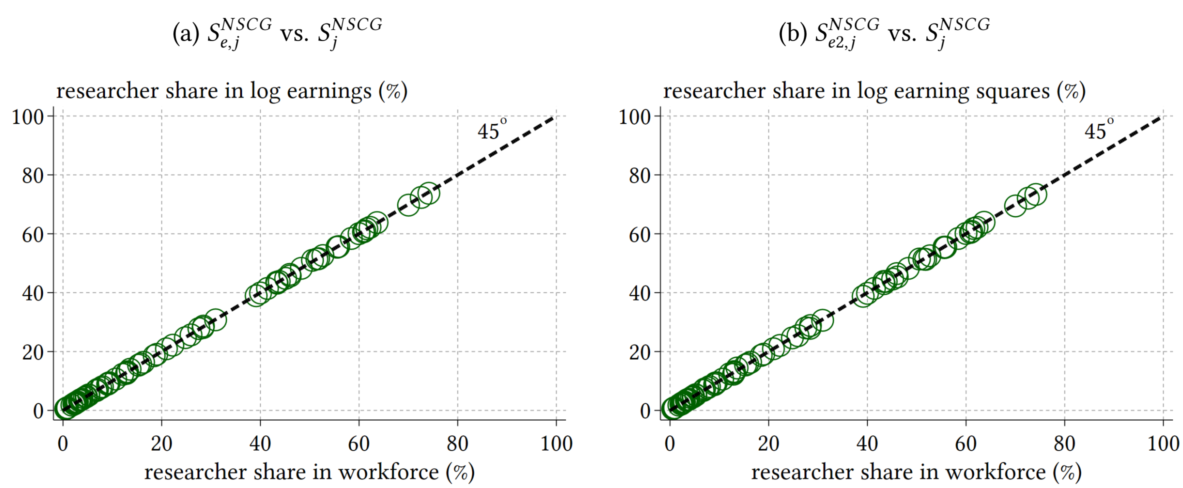
We also recalculate the trends of the sectoral mean and variance of constructed from the residual prooduced by a regression of log earnings in the ACS, controlling for age, age-squared, gender, and year-fixed effects. The resulting trends based on residualized log earnings are reported in A5 and A6; the former shows the mean log earnings of researcher and other college graduates relative to no-college workers; the latter shows the variance of the log earnings for the three sectors. We can see that all trends are robust to residualization.

Figure A2: Occupational researcher share (S_j^{NSCG}) in 1993 and 2003 against 2017



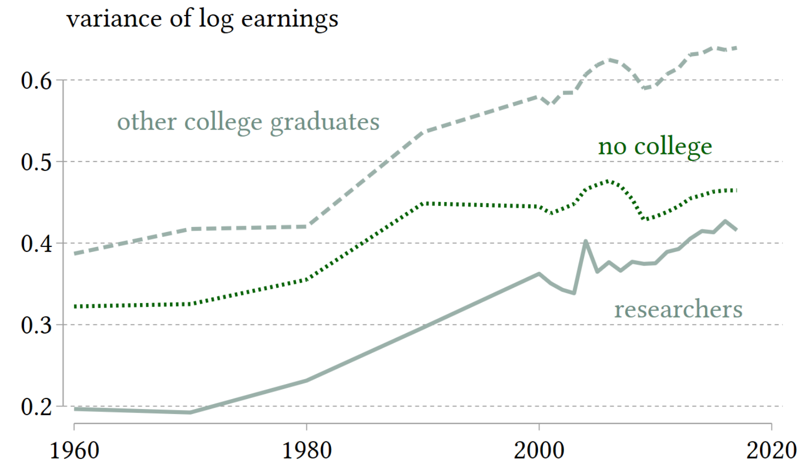
Notes: Each circle refers to an occupation in NSCG; the circle size is proportional to the number of workers in the occupation. Source: NSCG.

Figure A3: Occupational researcher share, log earnings and log earning squares



Notes: Panel (a) contrasts occupational researcher shares in log earnings ($S_{e,j}^{NSCG}$) against researcher shares in workforce (S_j^{NSCG}). Panel (b) contrasts occupational researcher shares in log earning squares ($S_{e2,j}^{NSCG}$) against researcher shares in the workforce (S_j^{NSCG}). Data source: NSCG.

Figure A4: Variance of log earnings, U.S. 1960-2017



Notes: Figure depicts the variance of log annual labor income for three types of labor—researcher, other college-educated workers, and workers without college degrees. Source: calculated by the authors from the NSCG and the Census/ACS.

F Supplementary materials for the NSCG panels

This section reports supplementary statistics for the earning distributions in the NSCG 2010-2013 and 2013-2015 panels, as used in Section 3.3. Figure A7 shows the analogous moments of 9 in the 2010-2013 panel. Table A1 reports the mean, standard deviation, and number of observations for each distribution drawn in the two figures. The results show that the patterns highlighted in Section 3.3 using the 2013-2015 panel also hold in the 2010-2013 panel.

Figure A8 shows that the primary findings produced from the NSCG panel are robust to using the residual log earnings controlling for age, age-squared, gender, race,and year-fixed effects. Panel (a)-(d) compares the earning distributions between movers and stayers using the 2010-2013 longitudinal data; Panel (e)-(h) makes the same comparison using the 2013-2015 longitudinal data. The main conclusions in Section 3.3 holds. Stayers earn more than movers in the research sector, while movers earn more than stayers in the non-research college sector. Moreover, mean log earnings gaps between movers and stayers are similar with or without residualizing log earnings.

Figure A5: Average residualized log earnings, relative to no-college workers, U.S. 1960-2017.

Notes: Log earnings in ACS are residualized by OLS regression controlling for age, age-squared, gender, and year-fixed effects. Source: calculated by the authors from the NSCG and the Census/ACS.

Figure A6: Variance of residualized log earnings, U.S. 1960-2017.

Notes: Log earnings in ACS are residualized by OLS regression controlling for age, age-squared, gender, and year-fixed effects. Source: calculated by the authors from the NSCG and the Census/ACS.

Table A1: Summary statistics of log earnings from the NSCG panel

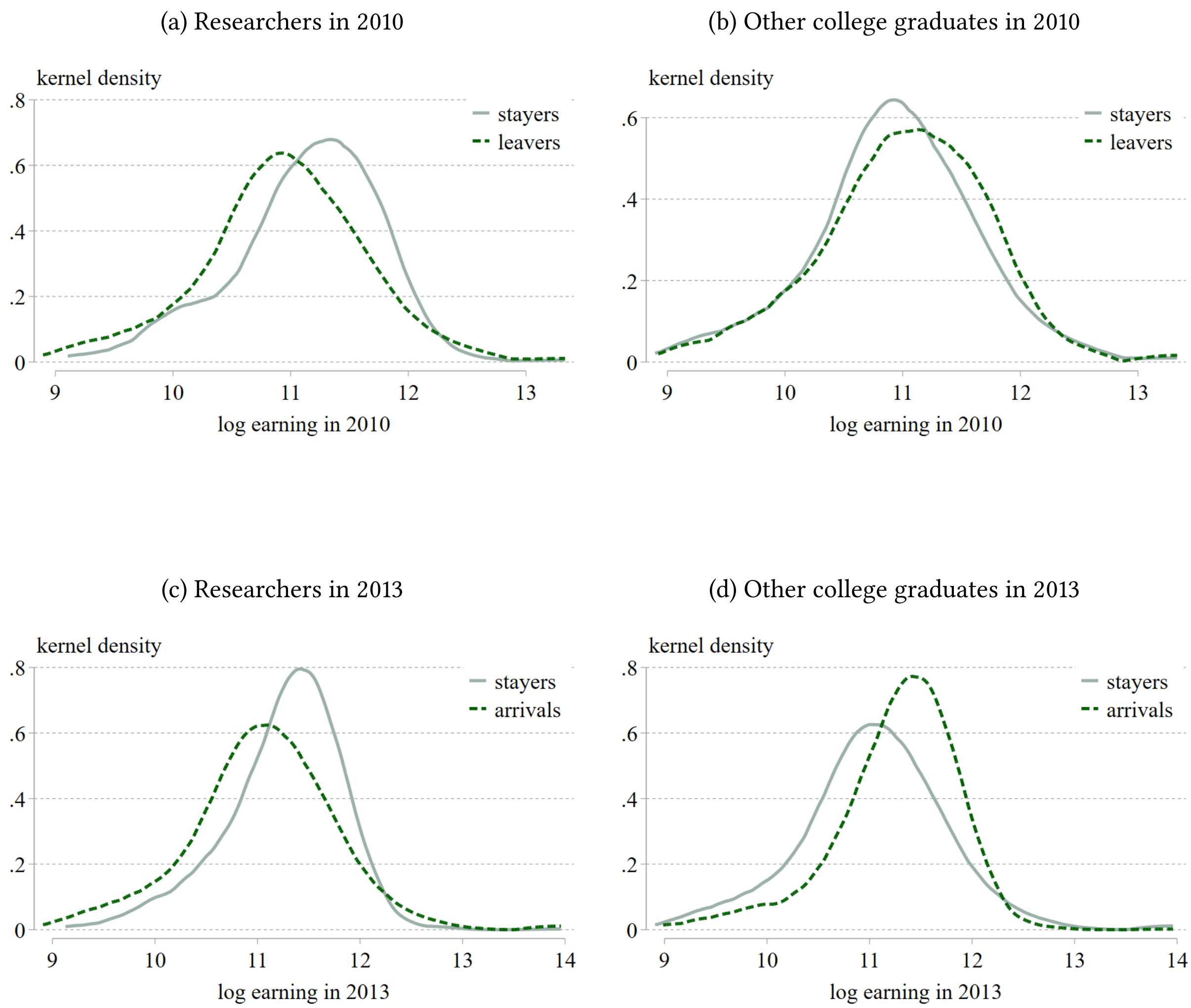
Panel (a): NSCG 2013-2015 panel

	log earnings in 2013				log earnings in 2015 panel			
	researchers (R)		other college (H)		researchers (R)		other college (H)	
	stayers	leavers	stayers	leavers	stayers	arrivals	stayers	arrivals
mean	11.10	11.07	10.96	11.09	11.24	11.16	11.07	11.25
std. dev.	0.673	0.716	0.744	0.688	0.598	0.636	0.738	0.610
observations	4,348	1,355	30,062	1,124	4,348	1,124	30,062	1,355

Panel (a): NSCG 2010-2013 panel

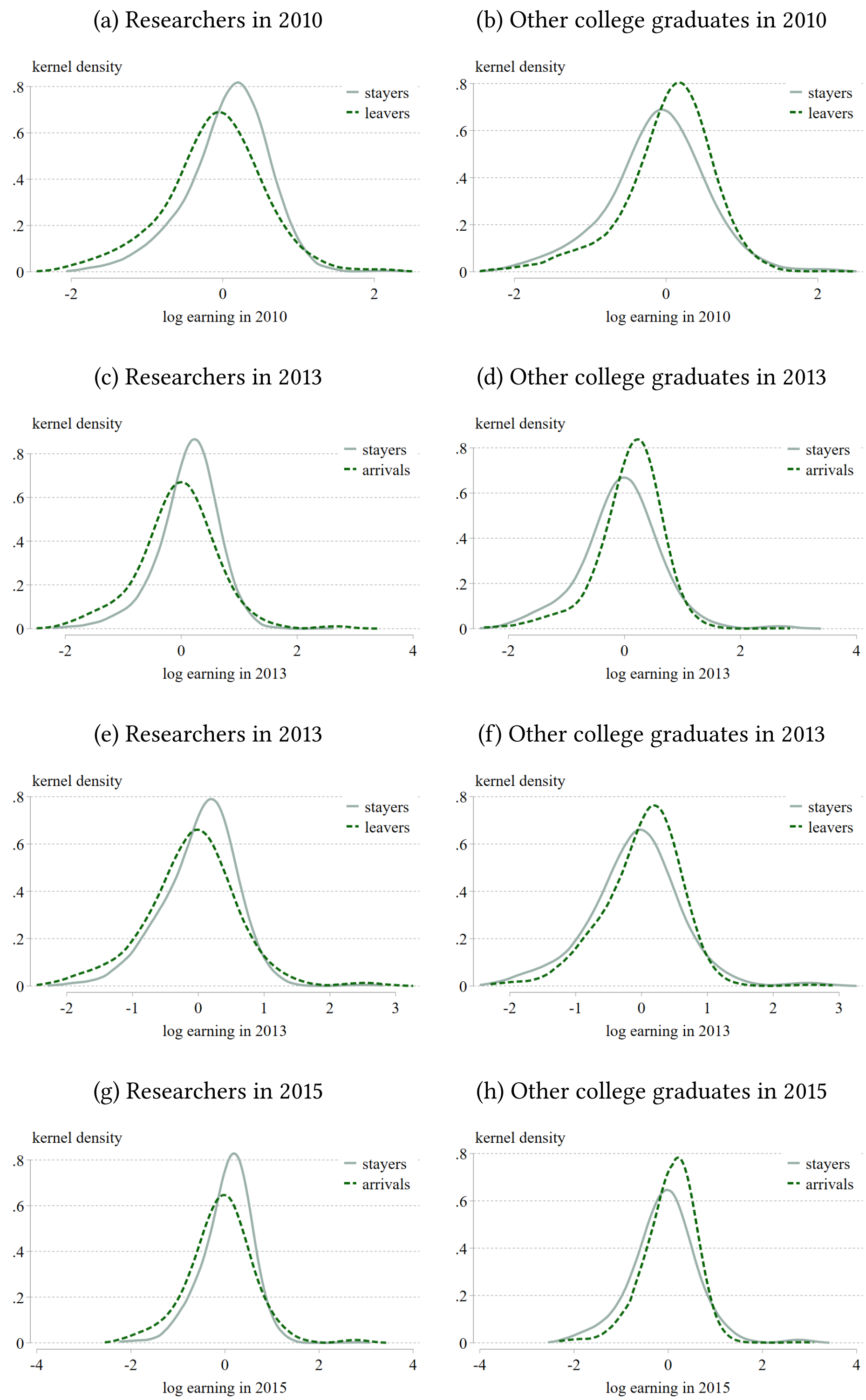
	log earnings in 2010				log earnings in 2013 panel			
	researchers (R)		other college (H)		researchers (R)		other college (H)	
	stayers	leavers	stayers	leavers	stayers	arrivals	stayers	arrivals
mean	11.16	11.01	10.95	11.21	11.28	11.25	11.06	11.21
std. dev.	0.612	0.724	0.690	0.675	0.552	0.594	0.708	0.663
observations	2,687	951	21,185	798	2,687	798	21185	951

Figure A7: Earning distributions in the NSCG 2010-2013 panel data



Notes: Data source: NSCG 2010-2013 panel.

Figure A8: Residualized earning distributions in the NSCG panel data



Notes: Data source: NSCG 2010-2015 panel. Panel (a)-(d) compares the earning distributions