On the Assignment of Workers to Occupations and the Human Capital of Countries*

Alexander Monge-Naranjo† Verónica Mies‡ Matías Tapia§

October, 2018

Abstract

A worker's human capital determines his absolute and relative productivity across different occupations. For a country as a whole, the cross-section distribution of workers determines the country's equilibrium assignment of workers to jobs and the resulting aggregate human capital. We consider a tractable general equilibrium Roy model and use it to infer, from observed data across countries and time: (i) the absolute and comparative advantage components of the different workers, (ii) the occupation intensity of a country's human capital, and (iii) the distortions in the allocation of workers to jobs. Contrary to the standard measure that implicitly assumes that human capital entails only absolute advantage, we show that the data imply that: (a) human capital has a strong comparative advantage component; (b) a higher distribution in the human capital distribution of workers leads to skill-upgrading across occupations and to a higher skill intensity of the overall human capital of countries; (c) cross-country differences in aggregate human capital explain a much larger fraction of the cross-country income differences from the standard model. We also find substantial costs from the distortions in the allocation of workers to jobs, especially for the less developed countries.

Keywords: Comparative Advantage; Factor Intensities; Cross-Country Income Differences; Misallocation.

JEL Codes: *J21, J24, J31, O15, O47.*

*We are thankful for comments and suggestions from Joe Kaboski, Todd Schoellman and seminar participants at Goethe University in Frankfurt and the 2018 SED conference in Mexico. Qiuhan Sun provided superb research assistance. *The views expressed here are those of the authors and do not necessarily reflect the opinion of the Federal Reserve Bank of St. Louis or the Federal Reserve System.*

†Federal Reserve Bank of St Louis and Washington University in St Louis

‡P. Universidad Católica de Chile

§Banco Central de Chile

1 Introduction

At any point in time, the labor markets of a country assign workers of different characteristics to multiple and alternative occupations and jobs. The workers' comparative advantages are the key factors underlying such an assignment, and with them are endogenously determined the relative supply and the valuation of the various skills used in production. Obviously, any notion of aggregate human capital for the country as whole must arise from the economy's equilibrium assignment of workers. Yet, while most economists would agree that comparative advantage is a key force allocating workers to occupations, most work on the importance of aggregate human capital for the income of countries ignores comparative advantages altoghether. The standard developmeent and growth accounting exercises simply add-up the schooling and other skills of workers, a treatment that is tantamount to impossing that the human capital of workers is an absolute advantage shifter across all occupations.[1]

This paper uses a simple general equilibrium Roy model to derive the aggregate human capital of countries. We show how the existence and uniqueness of an aggregate human capital arises from the general equilibrium assignment of the workers of an economy to the alternative occupations used in production. We then show how to use the equilibrium conditions from the model to infer from available data the underlying comparative and absolute advantage of the different workers, the production factor intensity of the different occupations in the country, the resulting skill prices, and the economy's aggregate human capital. Our framework also allows to assess the aggregate and distributional misallocation costs of wedges and taxes faced by the different workers in the different occupations. We then use our model to examine data on the observed assignment of workers to occupations at different points in time in the US states in different periods. We also examine data for a number of countries that span a large range of development levels. We find that human capital can explain a much larger share of the income differences of aggregate economies than what has traditionally been found, e.g. Caselli (2005) and Hanushek et al. (2017), supporting the notion that on top of an inferior supply of skills, poor countries are also characterized by skill unbalances in a sense that will be made precise below. We also find substantial misallocation of human capital in the poorer countries, supporting the notion that those poorer countries also tend to misuse their inferior supply of skills.

In our setting, a worker's human capital determines his absolute and relative productivity across different occupations. For a country as a whole, the cross-section distribution of workers and the factor intensity of the different occupations in production, determine the country's equilibrium assignment of workers to jobs and the resulting aggregate human capital. We consider a tractable general equilibrium Roy model that is embedded in an otherwise standard neoclassical framework. The standard measure commonly used in growth and development accounting is encompassed as the special case when human capital entails only absolute advantage. In our model,

[1] See for example Hall and Jones (1999), Caselli (2005), Hanushek et. al (2017), Hendricks and Schoelmann (2017), and many others.

workers' idiosyncratic abilities are jointly distributed according to an extreme value distribution, whose location parameters are determined by the human capital type of the worker across the different occupations. Production is modeled as a CES over the human capital levels used across all occupations, *not across different types of workers* as in Jones (2014) and Caselli and Ciccone (2018). Therefore, the aggregate human capital of a country is not only determined by the composition of its workers (who they are), or by what tasks are done in the country, (as emphasized in the structural transformation literature), but ultimately by the assignment of workers to occupations (i.e. who does what.) While in our framework the complementarity and/or substitutability of workers of different types is endogenously determined, and thus seemingly more sensitive to our specification of the economy, imposing the discipline of an equilibrium assignment provides a clean mapping between the economy's relative supply of workers of different types, the factor intensities of occupations in production with the resulting economywide aggregate human capital. Moreover, our framework naturally captures the impact of distortions to the assignment of workers and is very tractable to assess the aggregate costs of misallocations.

Using data on the observed assignment of workers to jobs, we show that human capital in the form of formal schooling has a strong comparative advantage component. For the country as whole, a higher distribution (in the first order sense) in the human capital distribution of workers leads to skill-upgrading across occupations, and to a higher skill intensity of the overall human capital of countries. When using our model with observed data on the assignments of workers to occupations, we find that the implied measure of aggregate human capital explains a substantially larger fraction of the income differences than the standard model. We find substantial costs from the distortions in the allocation of workers to jobs, especially for the less developed countries.

Section 2 describes the model environment, essentially a Roy model embedded in a neoclassical general equilibrium environment. The economy is populated by a finite (but arbitrarily large) set of 'types' of workers which can be allocated to a finite (but arbitrarily large) number of different occupations. The human capital 'type' of each worker determines their absolute advantage, their average efficiency units for all occupations uniformly. It also determines the worker's comparative advantage, a factor term that is worker-type-and-occupation specific, and hence, drives the relative equilibrium incidence of the different workers in each of the different occupations in the economy. We also allow for the possibility wedges or taxes that are worker-type-occupation specific and hence distort the equilibrium assignment in the economy. Finally, an idiosyncratic term that is distributed according to an extreme value distribution determines the probabilities with which each individual worker is assigned to the different occupations. This Roy setting is embedded in an otherwise standard neoclassical production economy, where the concept of human capital is extended to allow the provision of aggregate units of human capital to imperfectly substitutable –and possibly complementary– occupations. We show that the competitive equilibrium of this economy exists and is unique and gives rise to a well defined aggregate human capital. Moreover, the model has simple but useful comparative statics that can be directly mapped into the observed

patterns in the data on the allocation of workers to occupations and the relative factor shares or intensity of skills in output.

From observed data on the assignment of workers to occupations, Section 3 explains how the general equilibrium Roy model can be used to infer the underlying distribution of skills of the different human capital groups of workers in the economy. We first show how the conditions of undistorted equilibria can be used to deliver a simple characterization of the comparative advantage components simply from the observed allocation of workers of each type across all the occupations. The undistorted equilibrium also pins down the factor intensity shares of all occupations, the implied quantity and price for human capital in each occupation and the appropriate aggregate human capital. Second, we extend the analysis to distorted economies, where the presence of either taxes or wedges distort the allocation of workers to occupations. We show that when the economy is distorted by 'pure wedges', which we define as deadweight losses of productivity to both firms and workers, then the equilibrium implications are virtually unchanged, except that the comparative advantages terms must be corrected by the wedges. However, when the economy is distorted by 'pure taxes', which we define as the case when workers are not paid the productivity received by the firms, then a correction must be included in the formulas of the implied overall aggregate human capital and how it is distributed across occupations. Regardless of the form of the distortions, we show how to use the equilibrium model developed in Section 2 to conduct counterfactual experiments on removing the distortions, thus providing an assessment of the degree of misallocation in the economy.

According to our model framework, inferring the human capital of countries requires observing data on the allocation of workers to occupations. Admittedly, these heigthened data requirements reduce our ability to apply the model to the large set of countries included in recent development accounting exercises. In Section 4 we discuss the data available for our exercise. Our main source of data is the Integrated Public Use Microdata Series (IPUMS),[2] which collects individual level data for the U.S.A. and a large number of other countries. In this paper, we use the IPUMS data in two fronts. First, we collect data from IPUMS-USA on the U.S. states for a number of years, spanning from 1960 to 2016. Second, from IPUMS-International we collect data for 7 countries and for years that range from 1960 to 2000 and that span drastically different levels of development. We complement the IPUMS data with other standard sources, such as the Bureau of Labor Statistics (BLS), the Bureau of Economic Analysis (BEA) and the Penn World Table (PWT). Then, our quantitative exercises are divided in two groups. First, we use the model to assess the importance of human capital to account for the aggregate income differences across U.S. states. We do this exercise for different years, 1960, 1990, 2000 and 2016, and in all cases we find that human capital differences account for a substantially larger fraction of the state income differences. Second, we look at cross-country income differences for the handful of countries from which we have data. Here we find interesting differences in the contribution of human capital

[2] https://www.ipums.org/

differences relative to the standard, absolute advantage only model.

Our paper is closely related to the recent debate between B. Jones (2014) and Caselli-Ciccone (2018) about the usefulness of models with imperfect substitution of workers to explain income differences. On one hand, Jones (2014) argues that with imperfect substitution, the complementarity of higher skilled workers with the marginal product of low skilled workers will lead to an overstatement of the importance of TFP to explain output per worker differences. Caselli-Ciccone (2018) counters that this effect is either small or even negative. By looking directly at the assignment of workers to occupations, in this paper the attention is redirected to the substitution across occupations, not across workers types. Moreover, the issues of worker substitutability or complementarity are captured by the equilibrium of the Roy model, which we discipline by the occupation choices observed in the data.

Our Roy model draws on recent work by Acemoglu and Autor (2011), Costinot and Fogel (2010, 2014), Burstein et al. (2018) and Hsieh et al. (2018), which primary focus is on the equilibrium wage inequality arising in equilibrium as workers with different characteristics are allocated across occupations with different skills requirements. Our focus is instead on the country's aggregate effective supply of human capital and its contribution for aggregate output. In that sense, our focus is more closely related to that of Hsieh et al. (2018), who look at the aggregate impact of reducing misallocation driven by discrimination in education and labor markets in the U.S. However, our interest is in the role of cross-economy income differences. Moreover, our attention is in the relative supply differences of human capital types, specially education attainment groups, and not only on gender and race differences.

The reminder of the paper proceeds as follows. In the next section, we describe the model, characterize its equilibrium, proving its existence and uniqueness. Section 3 describes how we can use the equilibrium conditions of the model to infer, from observed data, the underlying absolute and comparative advantage and relative factor intensities. Section 4 describes the data. Section 5 describes the results for the U.S. states. Section 6 describes the international results. Section 7 concludes. Appendix A contains the proofs and Appendix B provides additional details on the data.

2 The Model

In this section, we describe the economic environment that we use to model the human capital of countries. We first lay out the environment and then define equilibrium and prove its existence and uniqueness. Then, we discuss how the aggregate human capital of a country is shaped from the endowment of workers and the technology of the country.

2.1 The Environment

As in the standard growth model, a composite final good Y_t is produced in each period t according to a production function $Y_t = Z_t F(K_t, H_t)$. Here, $F(\cdot,\cdot)$ is a constant returns to scale production function, Z_t is an exogenous Hicks-neutral productivity term (TFP), K_t is the flow of services from the stock of physical capital in the country, and H_t is the aggregate flow of human-capital augmented labor services. We extend the standard model by considering a human capital H_t that arise in equilibrium from the allocation of workers to tasks.

First, H_t is the resulting bundle of human capital services is given and is given by

$$H_t = G_t[H_t(1), ...H_t(J)],$$

where $G_t(\cdot)$ is a constant returns to scale function defined over the vector $H_t(j)$ of effective labor services provided in occupations or *'jobs'* j = 1,...,J in period t. Second, we consider the allocation of workers with different human capital '*groups*', indexed by $e = 1, .., E$ across the different occupations $j = 1, ..., J$. The aggregate supply levels $H_t(j)$ arise from optimal occupation choices, i.e. the general equilibrium assignments, of workers in groups e to occupations j, as explained below.

Much of our analysis can proceed with generic functions $F_t(\cdot)$ and $G_t(\cdot)$, but for concreteness we will focus on commonly used functional forms. Specifically, we adopt a Cobb-Douglas aggregate production function:

$$Y_t = Z_t (K_t)^{\alpha} (H_t)^{1-\alpha}, \tag{1}$$

where $0 < \alpha < 1$ is the physical capital share of output. Similarly, the human capital aggregator is given by a CES, i.e.

$$H_t = \left[\sum_{j=1}^{J} M_t(j) [H_t(j)]^{\rho}\right]^{\frac{1}{\rho}}, \tag{2}$$

where ρ is a parameter that indicates the degree of complementarity across the different occupations and can entertain values anywhere between $-\infty$, Leontieff, i.e. extreme complements, and $+1$, i.e.perfect substitutes.[3] Here, $M_t(j) \geq 0$ are the distributional factors that determine the share of human capital services in occupations j on the overall flow of human capital services H_t in the economy. We adopt the normalization $\sum_{j=1}^{J} M_t(j) = 1$ so that the TFP of the country Z_t captures all the Hicks neutral productivity shifts and $M_t(j)$ are the CES distributional parameters in the production function of H_t.

The population of workers inside the country in period t is described by a discrete distribution

$$S_t = [S_t(1), ..., S_t(e), ...S_t(E)].$$

Here, $S_t(e) \geq 0$ for all groups e, and we normalize the population measure to one, so that $\sum_{e=1}^{E} S_t(e) = 1$. Thus, $S_t(\cdot)$ is simply a discrete probability distribution, describing the cross-section of workers groups or types that populate the country at period t. In light of our data, we

[3] Values of ρ above 1 are ruled out to keep the aggregate H_t to be a concave function of $H_t(j)$.

will assume that the number of human capital groups E is finite. In our baseline exercises, we think of each e as indexing education attainment levels. We then extend our groupings of human capital to include work experience and gender.

As in the data, workers in all the different groups e could potentially provide labor services in any of the $j = 1, ..., J$ occupations. The human capital type e of a worker, however, determines the proclivity of those workers to choose the different occupations. In our model, the human capital e of a worker determines not only his absolute advantage in the different occupations, but also his comparative advantage relative to other workers.

Specifically, the assignment of workers to occupations is potentially driven by four factors: *(a)* the *unitary skill price* in each occupation, $w_t(j)$; it applies to all workers, regardless of their type e, entering in that occupation; *(b)* a productivity component $T_t(e, j) > 0$ that determines the *average potential productivity* of workers with human capital e in occupation j; and *(c)* a *random* component, $\eta(j)$, of the different workers for each possible occupation j; and *(d)* distortions or *wedges* $D_t(e, j)$ such as barriers and compensatory variations that can be specific to the pairings j and e. We take factors *(b)*-*(d)* as exogenously given but unitary skill prices *(a)* will be determined endogenously by the equilibrium of the economy.

We follow Burstein et. al (2016), Hsieh et. al (2016) and others, by assuming that each worker draws a random $1 \times J$ vector,

$$\eta = [\eta(1), .., \eta(J)] \in \mathbb{R}_+^J,$$

from a continuous joint distribution described by a p.d.f. $Q(\eta)$, where each $\eta(j)$ is drawn identically and independently, both, *(i)* across all individuals, and *(ii)* across all occupations for the same individual. In particular, we assume that the distribution is given by a multidimensional Frechet distribution:

$$Q(\eta) = \prod_{j=1}^{J} \exp\left\{-[\eta(j)]^{-\theta}\right\},$$

where $\theta > 1$ is a dispersion parameter that drives the degree of comparative advantage in the economy. We use a multidimensional Frechet distribution, a form of extreme value distribution, to exploit of a number of substantial analytical aspects in our quantitative analysis. In our formulation, $T_t(e, j)$ governs the abolute and comparative advantage of workers in group e for the provision of skills in occupations or jobs j. Indeed, the expected amount of labor units that each of the workers with human capital e supplies to occupation j is equal to $T_t(e, j)\Gamma\left(1 - \theta^{-1}\right)$ where $\Gamma(\cdot)$ is the Gamma function.[4]

2.2 Equilibrium Assignment

We consider competitive equilibria where the equilibrium assignment of workers to jobs is driven by the income maximization of workers, the profit maximization of firms and the clearing of all labor markets. The price system is simply a vector of unitary prices for each skill j, $w_t(j)$, and the

[4] That is, $\Gamma\left(1 - \theta^{-1}\right) = \int_0^\infty x^{-(1/\theta)} e^{-x} dx$.

rental rate of physical capital, denominated R_t, all in units of the final good. Taking those prices as given, workers of all groups e will be assigned to occupations j, generating a matrix $p_t(e,j)$ that indicates the fraction of workers in group e that work in occupation j. Here, $p_t(e,j) \geq 0$ and $\sum_{j=1}^{J} p_t(e,j) = 1$ for all e. Finally, taking as given the vector of unitary *occupation* or *skill* prices $w_t(j)$, the implied aggregate supply $H_t(j)$ must be equal to the demand of those human capital services from all the firms in this economy for the markets to clear.

The optimization conditions for firms and workers are straightforward. First, consider the firms. Because of constant returns to scale the firm size distribution is not pinned-down. Yet, any firms' hiring of services from the different forms of human capital services $H_t(j)$ and of physical capital R_t can be characterized by a stand-in firm that maximizes profits taking $w_t(j)$ and R_t as given, i.e.:

$$\max_{\{H_t(j),\, K_t\}} \left\{ Z_t (K_t)^{\alpha} \left(\left[\sum_{j=1}^{J} M_t(j) \left[H_t(j) \right]^{\rho} \right]^{\frac{1}{\rho}} \right)^{1-\alpha} - \sum_{j=1}^{J} w_t(j) H_t(j) - R_t K_t \right\}.$$

As expected, the first order conditions simply equate the wages $w_t(j)$ to the marginal products of the different forms of human capital, $H_t(j)$, i.e.:

$$w_t(j) = \bar{w}_t \times M_t(j) \left[H_t(j) \right]^{\rho - 1}, \tag{3}$$

where $\bar{w}_t \equiv (1-\alpha) Z_t (K_t/H_t)^{\alpha} \times (H_t)^{1-\rho}$ is an economywide component that is common across in the price of all skills j.

Second, consider workers. Each worker realizes a vector η and chooses the occupation that maximizes his net income,

$$\max_{i \in \{1,\ldots,J\}} \left\{ \eta(i) T_t(e,i) \frac{w_t(i)}{D_t(e,i)} \right\}.$$

This is a worker may opt for an occupation either because his particular human capital e usual leads to a *relatively* high value for the average $T_t(e,j)$ or because he happened to get a *relatively* high realization $\eta_t(j)$. However, with many ex-ante identical workers in each group e, the share of workers of the group that opts to each of the occupations j is equal to the ex-ante probability that any of them chooses them. Under the Frechet distribution, such probability is given by

$$p_t(e,j) = \frac{\left[w_t(j) \frac{T_t(e,j)}{D_t(e,j)} \right]^{\theta}}{\sum_{i=1}^{J} \left[w_t(i) \frac{T_t(e,i)}{D_t(e,i)} \right]^{\theta}}. \tag{4}$$

In what follows, we examine the equilibrium skill prices $w_t(j)$ and the allocations $p_t(e,j)$ that arise in equilibrium, given the cross-section of workers S_t, their productivity across occupations T_t and Q, the productivity shifts M_t and the wedges D_t. The equilibrium outcomes also determine the value of human capital levels $H_t(j)$ and H_t

2.2.1 Undistorted Equilibrium

In an undistorted equilibrium, $D(e,j)=1$, for all e and j. All workers get the same wage $w_t(j)$ for each unit of effective labor provided in occupation j. The assignment condition (4), becomes,

$$p_t(e,j)=\frac{\left[w_t(j)\,T_t(e,j)\right]^{\theta}}{\sum_{i=1}^{J}\left[w_t(i)\,T_t(e,i)\right]^{\theta}}. \tag{5}$$

Given the fractions $\{S_t(e)\}_{e=1}^{E}$, the total mass of workers in from education e in occupation j will be given by $q_t(e,j)=S_t(e)\,p_t(e,j)$ and the total fraction of individuals in occupation j in the country will be $q_t(j)=\sum_{e=1}^{E}S_t(e)\,p_t(e,j)$. More importantly, in terms of effective labor units, the total labor from workers with human capital e supplied to occupation j, $H_t(j,e)$, is given by

$$H_t(j,e)=\left[S_t(e)\,p_t(e,j)\right]\left[\Gamma\left(1-\theta^{-1}\right)T_t(e,j)\,p_t(e,j)^{-1/\theta}\right], \tag{6}$$

where the first term in brackets is the total mass of workers with human capital e working in occupations j and the second term is the average effective labor units provided by the group. Notice that this average goes down with the probability of entry, as the marginal worker entering has, on average, lower skills realizations $\eta(j)$. Summing over all human capital types e:

$$H_t(j)=\Gamma\left(1-\theta^{-1}\right)\sum_{e=1}^{E}S_t(e)\left[p_t(e,j)\right]^{(\theta-1)/\theta}T_t(e,j). \tag{7}$$

From here, we can write the total human capital H_t in the economy as

$$H_t=\Gamma\left(1-\theta^{-1}\right)\left[\sum_{j=1}^{J}M_t(j)\left[\sum_{e=1}^{E}S_t(e)\left[p_t(e,j)\right]^{(\theta-1)/\theta}T_t(e,j)\right]^{\rho}\right]^{\frac{1}{\rho}}.$$

After plugging expression (4) for $p_t(e,j)$, the aggregate human capital H_t becomes

$$H_t=\Gamma\left(1-\theta^{-1}\right)\left[\sum_{j=1}^{J}M_t(j)\left[\sum_{e=1}^{E}\frac{S_t(e)\,w_t(j)^{\theta-1}\,T_t(e,j)^{\theta}}{\left[\sum_{i=1}^{J}\left[w_t(i)\,T_t(e,i)\right]^{\theta}\right]^{(\theta-1)/\theta}}\right]^{\rho}\right]^{\frac{1}{\rho}},$$

which is driven by the cross-section $S_t(e)$, the average productivities $T_t(i,e)$, and the occupation productivities $M_t(j)$, but also by the given wages $w_t(j)$.

Plugging the first order condition for wages (3) into (5) and taking out the common factor $\bar{w}_t$, we obtain

$$p_t(e,j)=\frac{\left[M_t(j)\left[H_t(j)\right]^{\rho-1}T_t(e,j)\right]^{\theta}}{\sum_{i=1}^{J}\left[M_t(i)\left[H_t(i)\right]^{\rho-1}T_t(e,i)\right]^{\theta}}. \tag{8}$$

Next, plug this expression (8) into the formula (7) for the human capital services in occupation j to obtain the fixed-point conditions

$$H_t(j)=\left\{\Gamma\left(1-\theta^{-1}\right)\sum_{e=1}^{E}\frac{S_t(e)\,M_t(j)^{(\theta-1)}\,T_t(e,j)^{\theta}}{\left[\sum_{i=1}^{J}\left(M_t(i)\left[H_t(i)\right]^{\rho-1}T_t(e,i)\right)^{\theta}\right]^{(\theta-1)/\theta}}\right\}^{\frac{1}{1-(\rho-1)(\theta-1)}}. \tag{9}$$

Closed form solutions can be obtained when $\rho = 1$ and the different forms of human capital services $H_t(j)$ are perfect substitutes with each other. In this case, $w_t(j) = \bar{w}_t \times M_t(j)$ and the shares of workers e into occupations j are given by

$$p_t(e,j) = \frac{[M_t(j)\, T_t(e,j)]^{\theta}}{\sum_{i=1}^{J} [M_t(i)\, T_t(e,i)]^{\theta}}, \tag{10}$$

and the implied aggregate human capital levels per occupation, $H_t(j)$, are given by

$$H_t(j) = \Gamma\left(1-\theta^{-1}\right) \sum_{e=1}^{E} S_t(e) \frac{[M_t(j)]^{(\theta-1)} [T_t(e,j)]^{\theta}}{\left(\sum_{i=1}^{J} [M_t(i)\, T_t(e,i)]^{\theta}\right)^{(\theta-1)/\theta}}. \tag{11}$$

For the general case of $\rho \leq 1$, we have this simple but general result:

Proposition 1 *Consider an economy with a given configuration of workers $S_t(e)$, technological output shares $M_t(j)$, and average skills of workers e in jobs j, $T_t(e,j)$, with shape parameter $\theta > 1$. Then, for all $\rho \leq 1$ there exists a unique vector of occupation human capital levels $H_t(j)$ that solves the undistorted equilibrium fixed point condition (9.)*

As the others, the proof of this proposition is in the appendix. In any event, given the solutions to this fixed point problem in $\{H_t(j)\}_{j=1}^{J}$, we could readily compute the value of H_t, the equilibrium wages $w_t(j)$, and, of course the assignment of workers $p_t(e,j)$. Computing an equilibrium is straightforward to implement, which facilitates our quantitative exercises below.

2.2.2 Distorted Equilibria

We now consider the case in which different workers may receive different compensations for the same units of skills. To this end, we allow for wedges $D_t(e,j) \geq 1$ that reduce the effective supply of services of workers of type e into occupations j. With those wedges, the effective wages for workers in group e operating in occupation j are scale down by $1/D_t(e,j)$, i.e.

$$w_t(e,j) = w_t(j)/D_t(e,j),$$

where $w_t(j)$ is the marginal product to aggregate human capital services $H_t(j)$.

In all our analysis, we take the wedges as exogenously given, and consider two different formulations. In the first case, which we call pure wedges, the distortions $D_t(e,j)$ reduces the income of workers and their of effective supply of skills in equal measure. In the second case, the distortions $D_t(e,j)$ only reduce the earnings or utility of the workers, not his or her supply of skills. From the point of view of an individual workers, they are equivalent, but the two cases are different enough for the general equilibrium to merit a separate treatment.

$D_t(e,j)$ **as Pure Wedges** Besides scaling by $1/D_t(e,j)$ the earnings of workers e in occupations j, a wedge can also impact in the same proportion the effective supply of labor services received by the firm. The equilibrium supply of skills from group e is

$$H_t(e,j) = \Gamma\left(1-\theta^{-1}\right)\frac{S_t(e)\left[p_t(e,j)\right]^{(\theta-1)/\theta}T_t(e,j)}{D_t(e,j)}. \tag{12}$$

The equilibrium conditions (3) and (4) remain valid. Therefore, the wedges $D_t(e,j)$ not only impact the effective supplies directly, but also indirectly by affecting $p_t(e,j)$.

It is evident that the analysis for the undistorted equilibrium carries through as long as the terms $T_t(e,j)$ are replaced for $T_t(e,j)/D_t(e,j)$. Then, as with the undistorted case, we can take out the common factor $\bar{w}_t$ from (4) and obtain

$$p_t(e,j) = \frac{\left[M_t(j)\left[H_t(j)\right]^{\rho-1}\frac{T_t(e,j)}{D_t(e,j)}\right]^{\theta}}{\sum_{i=1}^{J}\left[M_t(i)\left[H_t(i)\right]^{\rho-1}\frac{T_t(e,j)}{D_t(e,j)}\right]^{\theta}}. \tag{13}$$

Next, plug this expression into the formula for (7) for the human capital services in occupation j to obtain a very similar set of fixed-point conditions as above:

$$H_t(j) = \left\{\Gamma\left(1-\theta^{-1}\right)\sum_{e=1}^{E}\frac{S_t(e)M_t(j)^{(\theta-1)}\left[\frac{T_t(e,j)}{D_t(e,j)}\right]^{\theta}}{\left[\sum_{i=1}^{J}\left(M_t(i)\left[H_t(i)\right]^{\rho-1}\frac{T_t(e,j)}{D_t(e,j)}\right)^{\theta}\right]^{(\theta-1)/\theta}}\right\}^{\frac{1}{1-(\rho-1)(\theta-1)}}.$$

Needless to say, when $D_t(e,j)$ are pure wedges, a distorted equilibrium with T_t is observable equivalent to an undistorted equilibrium with T_t/D_t. These wedges do not reflect discrimination or direct distortions in the labor market (as skills are paid their marginal product), but could reflect earlier forms of discrimination that occur when workers are acquring their skills, such as the quality of education received by women relative to men. In the quantitative exercises of the next sections, allowing for distortions $D_t(e,j)$ can be quite useful for the countries and years for which reliable wage data allows us to obtain an independent measure of the wedges.

$D_t(e,j)$ **as Implicit Taxes to Workers** As a variation, we can think of $D_t(e,j)$ as implicit taxes. In that case, even if there is not a reduction in the effective supply of services that each worker of type e delivers to the firms, the effective income for the workers is $w_t(e,j) = w_t(j)/D_t(e,j)$. This introduces a wedge between the contribution of the worker to the firm and the payoff he actually receives.

If wedges $D_t(e,j)$ distort the workers payoff but not their actual output, they can arise from different forms of distortions, e.g. discrimination, as analyzed by Hsieh et al. (2016). Wedges can be group specific, i.e. depending only on e, or occupation specific, i.e. depending only on j. Simple but extreme versions of wedges would be (a) a caste system in which for each group e, there

are pre-assigned occupations $\hat{J}(e)$ such $D_t(e,j)=1$ if $j\in\hat{J}(e)$ and $D_t(e,j)=\infty$ otherwise; or (b) an uniform barrier to only one occupation if $D_t(e,j^*)=\bar{D}>1$ for j^*. There wedges can reflect actual taxes, or utility costs associated to things like harrasment or poor working conditions. At any rate, any arbitrarily patterns across e and j can be considered in this setting.

Under this form of wedges, the solution fo $p_t(e,j)$ would be given by (13), but the expression for $H_t(e,j)$ is not (12) but the undistorted one, (6). Plugging the distorted $p_t(e,j)$ into the undistorted (6) and summing over all e, the supply of human capital services into occupation j, we get that the equilibrium allocation of $H_t(j)$ must solve the fixed point conditions

$$H_t(j)=\left\{\Gamma\left(1-\theta^{-1}\right)\sum_{e=1}^{E}S_t(e)\frac{\left[\frac{M_t(j)}{D_t(e,j)}\right]^{(\theta-1)}\left[T_t(e,j)\right]^{\theta}}{\left[\sum_{i=1}^{J}\left[M_t(i)\left[H_t(i)\right]^{\rho-1}\frac{T_t(e,j)}{D_t(e,j)}\right]^{\theta}\right]^{(\theta-1)/\theta}}\right\}^{\frac{1}{1-(\rho-1)(\theta-1)}}.$$

With the solution to $H_t(j)$, we compute the equilibrium value of H_t using (2), wages $w_t(j)$ using (3) and the equilibrium $p_t(e,j)$ using (4.)

2.3 A Convenient Decomposition

It is convenient to separate the role of pure absolute advantages across all occupations from those from comparative advantages for specific occupations. Consider a decomposition in the form $T_t(e,j)$ as $T_t(e,j)=A_t(e)C_t(e,j)$ where $A_t(e)$ is *uniform absolute productivity term of group* e across all *across all occupations* j and $C_t(e,j)$ is the comparative advantage term.[5] For brevity, we will discuss the pure wedges case only here.

Under such a decomposition, the assignment of workers becomes

$$p_t(e,j)=\frac{\left[w_t(j)\frac{C_t(e,j)}{D_t(e,j)}\right]^{\theta}}{\sum_{i=1}^{J}\left[w_t(i)\frac{C_t(e,i)}{D_t(e,i)}\right]^{\theta}}.$$

Two immediate implications follow. First, given wages, pure absolute advantage terms $A_t(e)$ do not affect the allocation of workers across occupations. Second, the general equilibrium determination of the wages $w_t(j)$ is how the aggregate equilibrium determines the assignment of different workers e across occupations j.

Similarly, the formula for the country's overall human capital services becomes:

$$H_t=\Gamma\left(1-\theta^{-1}\right)\left[\sum_{j=1}^{J}M_t(j)\left[\sum_{e=1}^{E}S_t(e)A_t(e)\frac{w_t(j)^{\theta-1}\left[\frac{C_t(e,j)}{D_t(e,j)}\right]^{\theta}}{\left[\sum_{i=1}^{J}\left[w_t(i)\frac{C_t(e,i)}{D_t(e,i)}\right]^{\theta}\right]^{(\theta-1)/\theta}}\right]^{\rho}\right]^{\frac{1}{\rho}},$$

[5] Any productivity shifts specific to an occupation but commong to all groups e would be captured by the terms $M_t(j)$, and, because of rescaling, in the TFP term Z_t.

which is strictly increasing in $A_t(\cdot)$ and homogeneous of degree one in $A_t(\cdot)$ and $S_t(\cdot)$. Solving for the general equilibrium determination of wages, and human capital across occupations, the fixed point that solves the equilibrium allocations becomes

$$H_t(j) = \left\{ \Gamma\left(1-\theta^{-1}\right) \sum_{e=1}^{E} S_t(e) A_t(e) \frac{M_t(j)^{(\theta-1)} \left[C_t(e,j)/D_t(e,j)\right]^{\theta}}{\left[\sum_{i=1}^{J} \left[M_t(i) \left[H_t(i)\right]^{\rho-1} \left[\frac{C_t(e,i)}{D_t(e,i)}\right]\right]^{\theta}\right]^{(\theta-1)/\theta}} \right\}^{\frac{1}{1-(\rho-1)(\theta-1)}}.$$

2.4 A Simple Benchmark: Absolute Advantage Only

Before characterizing the equilibrium assignment of workers to occupations of the model, it is convenient to consider the underlying, simpler case of absolute advantage only. Such a case boils down to the aggregate efficiency human capital units underlying in most the standard growth- and development-accounting analyses in the literature.

Assume **absolute but no comparative advantage**, across human capital types, i.e., for some $A_t(e) \geq 0$, we can write $T_t(e,j) = A_t(e)$ for all e and j. That is, the effective units of labor that a worker can provide is shifted *uniformly across all occupations* by the absolute advange term $A_t(e)$.

For now, consider a random and uniform assignment in which, regardless their human capital type e workers are randomly allocated to each occupation $j = 1, ..., J$ with probability $p(j)$, where $\sum p(j) = 1$. Then, the total human capital services $H_t(j)$ provided by group e to occupation j is simply $p(j) S_t(e) \Gamma\left(1-\theta^{-1}\right) B_t(j) A_t(e)$. Summing over all groups e, we obtain

$$H_t(j) = \Gamma\left(1-\theta^{-1}\right) p(j) \sum_{e=1}^{E} S_t(e) A_t(e).$$

Plugging this in the aggregator (2), the combination of all $H_t(j)$ leads to an aggregate human capital for the country in the form

$$H_t = \Gamma\left(1-\theta^{-1}\right) \left(\sum_{j=1}^{J} M_t(j) \left[p(j)\right]^{\rho}\right)^{\frac{1}{\rho}} \left(\sum_{e=1}^{E} S_t(e) A_t(e)\right),$$

The term inside brackets becomes completely indistinguishable from any other underlying the TFP term Z_t in the aggregate production function (1). Accounting for human capital boils down to the traditional measurement of the population's distribution of human capital levels, $S_t(e)$, and the absolute enhancements on productivity, $A_t(e)$, of the different types of human capital, as traditionally done using Mincer estimates.

Notice that this separation is indepedent of the particular choice of $p(j)$.[6] More interestingly, this form of indeterminacy is much more general. As as long as human capital only shifts absolute

[6] With some simple algebra it can be shown that, under the two conditions stated above, the optimal uniform

advantage, i.e. $T_t(e,j) = A_t(e)$, the term $\sum_{e=1}^{E} S_t(e) A_t(e)$ enters multiplicatively in the production function. The equilibrium assignment of human capital types e to occupations j is undetermined, and many different configurations of $p_t(e,j)$ would deliver the same $H_t(j)$ and H_t levels. The details for this argument are in the appendix. In any case, if human capital levels e only drive absolute advantage across all occupations, observing data on the allocation of those human capital groups across occupations, $p_t(e,j)$ would be uninformative about the human capital H_t of a country and the aggregate return of expanding the human capital endowments of countries, $S_t(\cdot)$ across the different levels e. This irrelevance result is overturned once we look into economies where comparative advantage drives the allocation of workers.

3 Inference from Observed Data

In this section we show how we can use the general equilibrium conditions of the model to infer the underlying distribution of skills of the different human capital groups. We first show how the conditions of undistorted equilibria can be used to deliver a simple characterization of the comparative advantage components $C_t(e,j)$ from the observed allocation $p_t(e,j)$. Next, we show how to extend the inference to include both the components of $T_t(e,j)$ and the wedges $D_t(e,j)$ when, besides the assignment data $p_t(e,j)$, we also observe earnings data $y_t(j,e)$. We should how to use these inferred values and general equilibrium assigment conditions to infer the aggregate human capital of countries.

3.1 Undistorted Equilibrium: Inferring $T_t(e,j)$ and $M_t(j)$

From the equation (5), we get that, for any two occupations j,j' and human capital groups, e, e', we can define a "ratio of ratios"

$$\left[\frac{p_t(e,j)/p_t(e,j')}{p_t(e',j)/p_t(e',j')}\right] = \left[\frac{T_t(e,j)/T_t(j',e)}{T_t(e,j)/T_t(j',e')}\right]^{\theta}. \tag{14}$$

This simple condition leads to three very useful implications about $A_t(e)$, $C_t(e,j)$ and $M_t(j)$. First, the comparative advantage terms $C_t(e,j)$ are proportional to $p_t(e,j)^{\frac{1}{\theta}}$. Therefore, only the distribution parameter θ is needed to infer the underlying comparative advantage terms from observed assignment data $p_t(e,j)$. Second, the occupation shifters $M_t(j)$ are direct drivers of $w_t(j)$ and therefore, can be inferred from the allocation of total human capital across occupations. Third, since absolute productivity components $A_t(e)$ have no bearing on the allocation of workers across

allocation $p^*(\cdot)$ across occupations is simply

$$p^*(i) = \frac{\left[T_t^J(i)\right]^{\frac{\rho}{1-\rho}}}{\sum_{j=1}^{J}\left[T_t^J(j)\right]^{\frac{\rho}{1-\rho}}},$$

for all $i = 1, ..., J$.

occupations, then data on $p_t(e,j)$ provides no information regarding $A_t(e)$. Absolute advantage can not be recovered from allocations, but inferred directly from income data, i.e. $A_t(e) = y_t(e)$. Formally we state the following:

Proposition 2 *Adopt the decomposition $T_t(e,j) = A_t(e) C_t(e,j)$ explained above and let $p_t(e,j)$ be the observed assignment of workers of human capital types $e = 1, ..., E$ into jobs $j = 1, ..., J$, and $A_t(e)$ an estimate of the absolute productivities of workers in group e. Then, if the underlying equilibrium of the economy is undistorted:* ***(a)*** *the comparative advantage term is given by*

$$C_t(e,j) = \bar{C}_t p_t(e,j)^{\frac{1}{\theta}},$$

for some positive $\bar{C}_t$ uniform across e and j; ***(b)*** *the pure relative occupation productivity terms $M_t(j)$ are given by*

$$M_t(j) = \frac{\left[\sum_{e=1}^{E} S_t(e) A_t(e) p_t(e,j)\right]^{1-\rho}}{\sum_{i=1}^{J}\left[\sum_{e=1}^{E} S_t(e) A_t(e) p_t(e,i)\right]^{1-\rho}}.$$

Proof. See Appendix. ■

From these simple results, we can obtain a very straightforward characterization of the aggregate human capital of a country.

Proposition 3 *Let $p_t(e,j)$ be the observed assignment of workers of human capital types $e = 1, ..., E$ into jobs $j = 1, ..., J$, and $A_t(e)$ an estimate of the absolute productivities of workers in group e. If the underlying equilibrium of the economy is undistorted, then the aggregate human capital of the country is given by*

$$H_t = \Gamma\left(1-\theta^{-1}\right) \frac{\left[\sum_{e=1}^{E} S_t(e) A_t(e)\right]^{\frac{1}{\rho}}}{\left\{\sum_{i=1}^{J}\left[\sum_{e=1}^{E} S_t(e) A_t(e) p_t(e,i)\right]^{1-\rho}\right\}^{\frac{1}{\rho}}}.$$

Proof. See Appendix. ■

An important result is that the resulting aggregate human capital is independent of the comparative advantage parameter θ. The key parameter is ρ, which governs the degree of substitution between occupations. The data on $p_t(e,j)$ determines the value of H_t unless $\rho = 1$ and occupations are perfect substitutes. If so, $H_t = \sum_{e=1}^{E} S_t(e) A_t(e)$, as in traditional measurements. Finally, as expected, notice that the equilibrium assignment of workers into occupations results in a country's that is increasing and homogeneous of degree 1 both in the quantity $S_t(e)$ of their workers as well as in their quality, $A_t(e)$.

3.2 Distorted Equilibrium: Inferring $\{T_t(e,j)\,,\, M_t(j)\,,\, D_t(e,j)\}$

In the previous section we showed how to use data on the assignment of workers to occupations, $p_t(e,j)$, and the average income $y_t(e)$ of each type of workers, and the conditions of an indistorted equilibrium to infer the comparative and absolute advantage components of the skills of workers, and the productivity parameters of each occupation in the aggregate human capital. In this section, we extend those inference exercises to economies in which frictions and/or compensating differentials can distort the assignment of workers to occupations. Such distortions generate income differences in conditional mean incomes each type e of wokers across occupations j.

First, we show how to perform a similar inference as the one before, for any form of wedges $D_t(e,j)$. Here, we show several simple results: (a), if the distortions are uniform across occupations j, i.e. $D_t(e,j) = \hat{D}_t(e)$ for all j, even if they vary across human capital groups e, they do not distort the allocation of workers. Those distortions only alter the absolute advantage of workers, and the inference of the previous section remain valid since they do not distort asignments; (b), if the conditional average income differences in the data are driven by pure wedges (in the sense defined above), the formula for aggregate human capital remains unchanged; (c) when the distortions are in the form of implicit taxes (in the sense defined above), the inferred values of aggregate human capital are affected, but simple formulas can still be obtained. In all those cases, changes in the wedges $D_t(e,j)$, would result in changes in the effective comparative advantage of workers, and in changes in the value of aggregate human capital of countries. Second, we show our simple method for using $y_t(e,j)$ data to infer wedges $D_t(e,j)$.

Inference of $C_t(e,j)\,, M_t(j)$ and H_t given $D_t(e,j)$: Consider having already gotten (inferred) data on $D_t(e,j)$ and on the average income $y_t(e)$ for workers of type e. First, consider that the wedges vary across workers, but not across occupations, i.e. $D_t(e,j) = \hat{D}_t(e) > 1$. In this case, the terms $\hat{D}_t(e)$ only distort the absolute advantage of workers. The true $A_t(e)$ underlying in the economy is the counterfactual income $y_t(e)\,\hat{D}_t(e)$ that group e would accrue in an undistorted economy. Other than that, the inferred comparative advantage terms $C_t(e,j)$ would not change.

Next consider the pure wedges case, where the average earnings $y_t(e,j)$ differences terms $D_t(e,j)$ reduce the effective supply of skills. In this case, the equilibrium assigment is given by the expression (4) leading to a ratio-of-ratios, for any pairs e, e' and j,j' of the form

$$\left[\frac{p_t(e,j)\,/p_t(e,j')}{p_t(e',j)\,/p_t(e',j')}\right] = \left[\frac{\frac{C_t(e,j)}{D_t(e,j)}/\frac{C_t(e,j')}{D_t(e,j')}}{\frac{C_t(e',j)}{D_t(e',j)}/\frac{C_t(e',j')}{D_t(e',j')}}\right]^{\theta}.$$

Therefore, the observed assignment of workers to occupations can be driven by either the underlying comparative advantages of workers of by the wedges they face in each occupation. From the previous expression, the only possible solution for $C_t(e,j)$ must necessarily be of the form

$$C_t(e,j) = \bar{C}_t D_t(e,j)\,[p_t(e,j)]^{\frac{1}{\theta}} \tag{15}$$

for any constant $\bar{C}_t$, which we normalize so as to the weight of the distortions are subsummed

into the absolute advantage $A_t(e)$, as discussed below. When plugging expression (15) into (4) the $D_t(e,j)$ terms cancel each other. Similarly, when plugging (15) into (6) those terms cancel out. Therefore, the same expressions as in the undistorted equilibrium for $H_t(e,j)$, $M_t(j)$, and H_t attain when $D_t(e,j)$ are pure wedges that reduce the effective supply of skills from workers.

Finally, consider the case where $D_t(e,j)$ are pure taxes that reduce the earnings of the worker but not the services received by firms. First, notice that the only solution for $C_t(e,j)$ is still (15) since occupational choices, from the point of view of the worker, are the same as the pure wedges case. Now, plugging $T_t(e,j) = A_t(e)\bar{C}_t D_t(e,j)[p_t(e,j)]^{\frac{1}{\theta}}$ into the undistorted expression $H_t(j,e) = [S_t(e)p_t(e,j)]\Gamma\left(1-\theta^{-1}\right)T_t(e,j)p_t(e,j)^{-1/\theta}$, normalizing $\bar{C}_t = 1$, simplifying and then summing over e, we obtain

$$H_t(j) = \Gamma\left(1-\theta^{-1}\right)\sum_{e=1}^{E} S_t(e)A_t(e)p_t(e,j)D_t(e,j). \tag{16}$$

That is, the effective aggregate supplies of skills j include the taxes $D_t(e,j)$ on workers. Since aggregate skill prices, $w_t(j) = \bar{w}_t \times M_t(j)[H_t(j)]^{\rho-1}$, are equalized, then

$$M_t(j)\left[\sum_{e=1}^{E} S_t(e)A_t(e)p_t(e,j)D_t(e,j)\right]^{\rho-1} = M_t(i)\left[\sum_{e=1}^{E} S_t(e)A_t(e)p_t(e,i)D_t(e,i)\right]^{\rho-1}.$$

Then, impossing $\sum_{i=1}^{J} M_t(i) = 1$ and solving,

$$M_t(j) = \frac{\left[\sum_{e=1}^{E} S_t(e)A_t(e)p_t(e,j)D_t(e,j)\right]^{1-\rho}}{\sum_{i=1}^{J}\left[\sum_{e=1}^{E} S_t(e)A_t(e)p_t(e,i)D_t(e,i)\right]^{1-\rho}}, \tag{17}$$

i.e. the tax distortions would enter in our inference for the distributional occupational weights $M_t(j)$.

4 Data

In this section, we describe our data on the assignment of workers to occupations for US states and a handful of countries with very different levels of development. We first describe our data sources, describing countries for which we have available data, describing the types of workers and occupations, as well as the different variables available to group the workers of those countries. Then, we describe how the assignment of workers to the different jobs varies across those country-year pairs.

4.1 US Data

All data are extracted from the Integrated Public Use Microdata Series, USA (IPUMS). Based on data availability, we perform a cross-section (50 states and Washington D.C.) time series (year

1960, 1990, 2000, and 2016) analysis. Data for year 1960, 1990, and 2000 contain 1-in-20 national random sample of the population while data for year 2016 contain 1-in-100 national random sample of the population.[7]

From IPUMS, we collect the following variables:

- age: Respondent's age in years
- sex: Respondent's gender
- statefip: The state in which the respondent was located
- educ: Respondent's educational attainment, as measured by the highest year of school or degree completed
- empstat: Indicates whether the respondent was a part of the labor force – working or seeking work – and, if so, whether the person was currently unemployed
- occ: Respondent's occupation, coded according to the 1950 Census Bureau occupational classification system. Occupation categories are:[8]
  - Managers
  - Professionals
  - Technicians
  - Clerks
  - Service workers
  - Agricultural workers
  - Traders
  - Operators
  - Elementary occupations
- incwage: Respondent's total pre-tax wage and salary income for the previous year

We only keep observations for employed individuals, given by empstat. To characterize occupations, we drop all individuals that do not fit in one of the nine occupation categories: Managers, Professionals, Technicians, Clerks, Services, Agriculture, Traders, Operators and Elementary. This means that we exclude individuals with other occupations, individuals with unknown or missing occupations, and individuals where the category is "Not yet classified".

For income, individuals with non-positive values, missing values, or top codes are eliminated from the database.[9] Additionally, we adjust income variables for inflation and express all figures in U.S. dollars of 2000.[10]

[7] More information on sample sizes can be obtained at https://usa.ipums.org/usa/sampdesc.shtml.

[8] More information on occupations can be obtained at https://usa.ipums.org/usa-action/variables/OCC1950#codes_section

[9] The top code for 1960, 1990, 2000, and 2016 is \$25,000, \$140,000, \$175,000, and 99.5th percentile in state, respectively. More information on top codes can be obtained at https://usa.ipums.org/usa-action/variables/INCWAGE#codes_section.

[10] Consumer price index adjustment factors for the appropriate years can be found in the CPI99 variable: https://usa.ipums.org/usa/cpi99.shtml. CPI99 provides the CPI-I multiplier available from the Bureau of Labor Statistics to convert dollar figures to constant 1999 dollars. This corresponds to the dollar amounts in the 2000 census, which inquired about income in 1999.

Education measures are derived from educ. We define seven education categories: (1) no schooling, (2) incomplete primary, (3) complete primary, (4) incomplete secondary, (5) complete secondary, (6) incomplete tertiary, and (7) complete tertiary. Individuals with missing observations are dropped.

For demographics, we define three age groups: "young" individuals aged between 25 and 35 years old, "middle aged" individuals between 36 and 50, and "old" individuals with ages above 50. Individuals below 25 or with missing data on age are dropped from the samples. Individuals with missing data on gender are also dropped.

4.2 International Data

IPUMS-International is also our source of cross-country comparisons. The records are converted into a consistent format that allows for cross-country comparisons. While the universe of international census records included in IPUMS is very large, the need for reliable data on earnings by occupations and worker's characteristics reduces our sample to only a handful of country-year pairs. Table 1 describes the countries and years for our we have data.

Table 1: Country-Years Available
IPUMS Data Available

1960s-1970s		**1990s-2000s**	
Country	**Year**	**Country**	**Year**
USA	*1960*	*USA*	*2000*
Canada	*1971*	*Canada*	*2001*
Indonesia	*1976*	*Indonesia*	*1995*
Mexico	*1960*	*Mexico*	*1990*
		Brazil	*2000*
		India	*1999*
		Panama	*2000*

For each of these country-year pairs, we collect individual level census data on age, gender, occupation, annual labor income (in 2000 USD PPP), and schooling attainment for all occupied workers 25 or older[11]. Occupations are directly comparable across countries, and are coded according to the major categories in the 1988 International Standard Classification of Occupations (ISCO) scheme.

All other variable definitions are similar to those for the U.S. data.

[11] Details are presented in the appendix

5 The Aggregate Human Capital of U.S. States

In this section, we adopt the basic results of the development accounting literature, e.g. Caselli (2005) and adapt it to account for the cross-state income per worker for the U.S. states. We follow Hanushek and consider only 47 states, excluding Alaska, Delaware, and, Wyoming, who argues that the incident of natural resources and/or tax policies, make them it less suitable for a neoclassical model and more difficult to compare to the other states. We also exclude the District of Columbia, since the behavior of employment to population and the incidence of government employment is substantially different from the remaining states.[12] In particular, we collect the following variables:

- income: Measured by gross domestic product (GDP) of private industries in millions of current dollars. Source: BEA
- capital: Measured by current-cost net stock of private fixed assets in millions of current dollars. Source: BEA
- employment: Measured by employees on nonfarm payrolls, seasonally adjusted, in thousands of people. Source: BLS Payroll Survey, Haver Analytics
- employment (IPUMS): The raw data count the number of employed people from IPUMS. The refined data exclude observations with missing occupation, education attainment, income level, gender, or age. Source: IPUMS-USA
- population (Source: Census Bureau, Haver Analytics): Measured by annual resident population in thousands of people. The U.S. resident population includes all persons who usually reside in the 50 states and the District of Columbia.[13]

We use data for 4 years: 1963, 1990, 2000, and 2016. In addition to the basic development accounting for each of the years, we also we perform two growth accounting exercises: 1960/3-1990 and 2000-2016. Doing so, we avoid a discontinuity in the series of capital and GDP in 1997.[14]

Figure 1 shows the result of the leading –and commonly used– success ratio proposed by Caselli (2005):

$$\text{success}=\frac{var\left[\log\left[(h_t^i)^{1-\alpha}\right]\right]}{\log\left[(y_t^i)\right]},$$

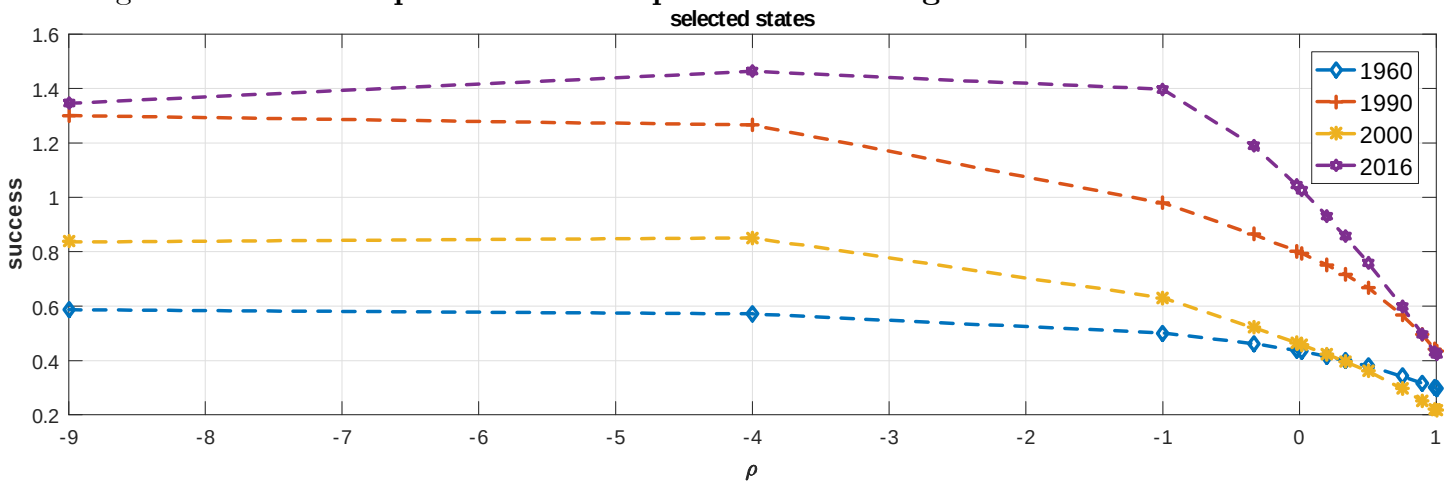
where h_t^i and y_t^i are the human capital and income per worker in each of the $i = 1, ..., 47$ U.S. states included in the sample for the periods $t = 1960, 1990, 2000, 2016$. We use the standard value

[12] Yet, our main results are not too sensitive to these exclusions.

[13] In addition, it excludes U.S. Armed Forces overseas and civilian U.S. citizens whose usual place of residence is outside the United States.

[14] There is a discontinuity in the time series of GDP by state at 1997, where the data change from SIC industry definitions to NAICS industry definitions. This discontinuity results from many sources. The NAICS-based statistics of GDP by state are consistent with U.S. gross domestic product (GDP) while the SIC-based statistics of GDP by state are consistent with U.S. gross domestic income (GDI). With the comprehensive revision of June 2014, the NAICS-based statistics of GDP by state incorporated significant improvements to more accurately portray the state economies. This data discontinuity may affect both the levels and the growth rates of GDP by state.

Figure 1: **Human Capital and Development Accounting for the U.S. States.**



of $\alpha = 2/3$. The figure shows the results for the different values of ρ.

There are three key results in this figure. First, there is quite a bit of variation in the importance of human capital in the accounting for income per worker differences across states. For instance, it is generally much larger in 2016 than in 1960, but smallr for 2000 than for 1990. However, given the results in the literature, it is quite reassuring, that these differences tend to collapse as the value of ρ tends to 1, i.e. the standard value that ignores comparative advantage. Second, the implied contribution of human capital when $\rho = 1$ is very similar to those obtain by Caselli (2005) for countries and by Hanushek et al. (2018) for the U.S. states using the standard measurement of aggregate human capital.

The third and most salient result is that moving away from the standard practice of ignoring comparative advantage as a driver of the assignment and aggregation of human capital would lead to potentially much higher explanatory power to human capital differences in explaining the income differences across US states. Indeed, using $\rho = 1/3$, a fairly conservative value as used by Hsieh et al (2018), would imply the success measure to approach 1 for 2016.

6 The Aggregate Human Capital of Countries

We take on the vast development accounting literature and re-examine the contribution of human capital differences. Since we compare economies at very different points of development, in this section we proceed with very conservative approach to avoid confounding human capital differences with TFP differences. For concreteness, denote the *country* i, *output per-worker* as

$$y_i = (Z_i)^{\frac{1}{1-\alpha}} \left(\frac{k_i}{y_i}\right)^{\frac{\alpha}{1-\alpha}} \times h_i.$$

As argued all along this paper, the usual approach of adding up the education of workers as absoluted advantage shifters can misclassify human capital into TFP differences. Let $h_i^{\text{pwt,9.0}}$ be the measure of human capital as computed in PWT 9.0. and let h_i the correct human capital per

worker according to our model, which is denoted in U.S. dollars of 2000, not in unskilled labor units of the country i. These two measures are not readily comparable. However, we can also use our model to compute $h_i^{\text{ps/aa}}$, the model's implied human capital when, $\rho = 1$, i.e. the model for the case of perfect substitutes/absolute advantage only. Thus, we can use the ratio $\frac{h_i}{h_i^{\text{ps/aa}}}$ to correct the PWT measure. Following these logic, the corrected income accounting equation should be:

$$y_i = \underbrace{\left((Z_i)^{\frac{1}{1-\alpha}}\right)}_{\text{TFP}} \times \underbrace{\left[\left(\frac{h_i}{h_i^{\text{ps/aa}}}\right)_{\text{model}} \times h_i^{\text{pwt,9.0}}\right]}_{\text{HK}} \times \left(\frac{k_i}{y_i}\right)^{\frac{\alpha}{1-\alpha}}.$$

Having corrected the measure of human capital, we can construct a simple success measure as:

$$\text{success} = \left[\frac{y_i^{(h_{\text{bechmark}})}/y_i^{\text{actual}}}{y_i^{\text{benchmark}}/y_i^{\text{actual}}}\right] = \frac{h_{\text{bechmark}}/h_i}{y_i^{\text{benchmark}}/y_i^{\text{actual}}},$$

where 'benchmark' indicates either the country at the end of the period (growth accounting) or the U.S. in 2000. The variable $y_i^{(h_{\text{bechmark}})}$ indicates the counterfactual output level if the country i was set to have the same human capital level as the benchmark economy.

We first use this equation to explore the contribution of human capital growth in explaining income per-capita growth for the fourth countries for which we can perform such a calculation. Table 2 shows the results for parameter $\rho = 1/3$, in the ballpark of the values used in Burstein et al. (2018) and Hsieh et al. (2018), and for the case of strong complementarities, $\rho = -2$.

Table 2: Human Capital and Growth Accounting

Country	**Years**	**Gross Growth**				**success H**		
		$h_i^{\text{pwt,9.0}}$	$h_i^{\text{corrected}}$		Y_{pw}	$h_i^{\text{pwt,9.0}}$	$h_i^{\text{corrected}}$	
			$\rho = 1/3$	$\rho = -2.$			$\rho = 1/3$	$\rho = -2.$
USA	*2000-1960*	1.32	1.33	1.30	3.39	0.39	0.39	0.38
Canada	*2001-1971*	1.25	1.31	1.43	1.39	0.89	0.94	1.03
Indonesia	*1995-1976*	1.40	1.46	1.17	2.18	0.64	0.67	0.54
Mexico	*1990-1960*	1.38	1.16	0.77	1.37	1.01	0.84	0.57

Notice that, contrary to the implications obtained for the US states, using the corrected measure of human capital may reduce not enhance the contribution of human capital for the growth of income of countries. Mexico is the most clear case. The standard PWT human capital measure, by itself, explains the growth in the output per worker in Mexico between 1960 and 1990. Such a contribution falls to only 84% using the baseline parameter value and falls much farther one we push the degree of complementarities. Interestingly, the opposite happens for Canada, where our corrected measure of human capital explains a larger share.

A similar pattern of changing the relative is observed when we use the correction to account for cross-country income differences, i.e. development accounting. Table 3 shows the success ratio of human capital in accounting for the gaps of the handful of countries in the sample for which we have data near 2000, and compare them to the U.S. income per worker as of 2000.

Table 3: Human Capital and Development Accounting

Country	Year	Relative to USA 2000				success H		
		$h_i^{\text{pwt,9.0}}$	$h_i^{\text{corrected}}$		Y_{pw}	$h_i^{\text{pwt,9.0}}$	$h_i^{\text{corrected}}$	
			$\rho = 1/3$	$\rho = -2.$			$\rho = 1/3$	$\rho = -2.$
Brazil	*2000*	0.57	0.52	0.49	0.21	0.37	0.40	0.43
Canada	*2001*	0.98	1.16	1.36	0.64	0.65	0.55	0.47
India	*1999*	0.48	0.49	0.58	0.05	0.11	0.11	0.09
Indonesia	*1995*	0.57	0.59	0.61	0.12	0.22	0.21	0.20
Mexico	*1990*	0.61	0.56	0.54	0.28	0.46	0.49	0.51
Panama	*2000*	0.72	0.71	0.84	0.26	0.36	0.37	0.31

Using the corrected measure of human capital increases the contribution of this factor for Brazil and Mexico, reduces it for Canada, and keeps it virtually unchanged for the rest.

We now explore the cost of misallocation. For all workers e and occupations j, we compute from the data the distortions for as Distortions are calculated as:

$$D(e,j) = \frac{w_{\max}(e)}{w(e,j)}.$$

Then, we follow our previous analysis and consider $D(e,j)$ either as 'pure wedges' or 'pure taxes'. Using the implied values for $A(e)$, $C(e,j)$ and $M(j)$ for each economy and case, we then do the counterfactual exercise of computing what would be the gains in the country's aggregate human capital of eliminating those distortions (setting $D(e,j) = 1$ for all) or just setting them to the level of the U.S.

Table 4: Misallocation Costs of Distortions D: % Gains in H.

Country	Year	Setting $D(e,j) = 1$		Setting $D(e,j) = D^{USA,2000}$	
		D: wedges	D: taxes	D: wedges	D: taxes
Brazil	*2000*	87.6	2.9	42.1	1.7
Canada	*2001*	24.4	2.0	-3.0	0.2
India	*1999*	49.8	3.9	9.9	2.5
Indonesia	*1995*	31.6	1.4	2.6	0.3
Mexico	*1990*	67.2	1.9	25.4	0.7
Panama	*2000*	56.3	4.2	19.6	2.6
USA	*2000*	28.5	2.0	–	–

The results are in Table 4. Not surprisingly, the costs of misallocation are substantially larger when the distortions are in the form of deadweight losses for both, firms and workers than when it is the form of pure taxes. Yet, it is interesting that they are one order of magnitude larger. In any event, our results indicate that poor countries such as Brazil, Mexico, India and Panama, seem to be highly distorted. A country such as Canada seem to be much less distorted. Indeed, it is less distorted than the U.S. and its human capital would fall if its distortions were set to the U.S. levels.

7 Conclusion

Using a simple general equilibrium Roy model, we show how the existence and uniqueness of an aggregate human capital arises from the general equilibrium assignment of the workers of an economy to the alternative occupations used in production. We then show how to use the equilibrium conditions from the model to infer from available data the underlying comparative and absolute advantage of the different workers, the production factor intensity of the different occupations in the country, the resulting skill prices, and the economy's aggregate human capital. Our framework also allows to assess the aggregate and distributional misallocation costs of wedges and taxes faced by the different workers in the different occupations. We then use our model to examine data on the observed assignment of workers to occupations at different points in time in the US states in different periods. We also examine data for a number of countries that span a large range of development levels. We find that human capital can explain a much larger share of the income differences of aggregate economies than what has traditionally been found, e.g. Caselli (2005) and Hanushek et al. (2017), supporting the notion that on top of an inferior supply of skills, poor countries are also characterized by skill unbalances in a sense that will be made precise below. We also find substantial misallocation of human capital in the poorer countries, supporting the notion that those poorer countries also tend to misuse their inferior supply of skills.

Using data on the observed assignment of workers to jobs, we show that human capital in the form of formal schooling has a strong comparative advantage component. For the country as whole, a higher distribution (in the first order sense) in the human capital distribution of workers leads to skill-upgrading across occupations, and to a higher skill intensity of the overall human capital of countries. When using our model with observed data on the assignments of workers to occupations, we find that the implied measure of aggregate human capital explains a substantially larger fraction of the income differences than the standard model. We find substantial costs from the distortions in the allocation of workers to jobs, especially for the less developed countries.

8 References

References

[1] Burstein, A., Morales, E. and Vogel, J. (2018) "Changes in Between Group Inequality: Computers, Occupations, and International Trade" American Economic Journal: Macroeconomics, forthcoming.

[2] Caselli, Francesco (2005). Accounting for Cross-Country Income Differences, in Philippe Aghion and Stephen Durlauf (eds.), Handbook of Economic Growth, Volume 1A, Elsevier.

[3] Caselli, Francesco (2017). Technology Differences over Space and Time, Princeton University Press

[4] Caselli, A. and Ciccone, A. (2018) The Human Capital Stock: A Generalized Approach, Comment. American Economic Review, forthcoming.

[5] Costinot, A. and Vogel, J. (2010). Matching and inequality in the world economy. Journal of Political Economy 118 (4), 747-786.

[6] Hall, Robert and Charles Jones, (1999). "Why Do Some Countries Produce So Much More Output Per Worker Than Others?" Quarterly Journal of Economics, 114, pp. 83-116.

[7] Hendricks, Lutz, and Todd Schoelmann (2017): Human Capital and Development Accounting: New Evidence from Wage Gains at Migration, Quarterly Journal of Economics.

[8] Hanushek, .E., Ruhose, J. and Woessmann, L.. (2017). American Economic Journal: Macroeconomics 2017, 9(4): 184–224.

[9] Hsieh, C-T., Hurst, E., Jones, C., Klenow, P. (2018) The Allocation of Talent and U.S. Economic Growth. NBER working paper, 18693.

[10] Jones, Benjamin (2014). The Human Capital Stock: A Generalized Approach, American Economic Review, 104(11), pp. 3752-3777.

A Proofs

Proof of Proposition 1. Take logs, and define $h_t(j) = \ln(H_t(j))$. Using expression (9) and for any $h \in \boldsymbol{R}^J$, define the mapping $\mathcal{T}(\cdot)$ as

$$(\mathcal{T}h_t)(j) = \ln\left[\left\{\Gamma\left(1-\theta^{-1}\right)\sum_{e=1}^{E}\frac{S_t(e)\,M_t(j)^{(\theta-1)}\,T_t(e,j)^{\theta}}{\left[\sum_{i=1}^{J}\left(M_t(i)\left[e^{h_t(i)}\right]^{\rho-1}T_t(e,i)\right)^{\theta}\right]^{(\theta-1)/\theta}}\right\}^{\frac{1}{1-(\rho-1)(\theta-1)}}\right]$$

It is easy to see that given $S_t(e)$, $M_t(j)$, $T_t(e,j)$, $\mathcal{T}$ maps vectors in the $\boldsymbol{R}^J$ into vectors in $\boldsymbol{R}^J$. Pick any continuous norm $||\cdot||$ in $\boldsymbol{R}^J$ and define the distance $d(h,g) = ||h-g||$. Obviously, $\left(\boldsymbol{R}^J, d\right)$ is a Banach space. To check that $\mathcal{T}$ is a contraction in $\left(\boldsymbol{R}^J, d\right)$, we simply verify that it satisfies Blackwell's sufficient conditions. To this end, consider any positive number $a > 0$. We have that $\mathcal{T}(h+a)$ is simply given by

$$\begin{aligned}\mathcal{T}(h+a)(j) &= \frac{1}{1-(\rho-1)(\theta-1)}\left\{\ln\left[\Gamma\left(1-\theta^{-1}\right)\right] + \ln\left(\sum_{e=1}^{E}\frac{S_t(e)\,T_t(e,j)\left[M_t(j)\,T_t(e,j)\right]^{(\theta-1)}}{\left(\sum_{i=1}^{J}\left[M_t(i)\left[e^{h(i)+a}\right]^{\rho-1}T_t(e,i)\right]^{\theta}\right)^{(\theta-1)/\theta}}\right)\right\} \\ &= \mathcal{T}(h)(j) + \frac{(\theta-1)(1-\rho)}{1+(1-\rho)(\theta-1)}a\end{aligned}$$

For $\mathcal{T}$ to be *monotone* it is sufficient that $(\theta-1)(1-\rho) > 0$, which, since $\theta > 1$ holds for all $\rho < 1$. Notice that $\rho < 1$ also suffices for *discounting.* This completes que proof that $\mathcal{T}$ is a contraction for $\rho < 1$. Moreover, since $H_t(j) = \exp(h_t(j))$, the implied solutions are valid since they are strictly positive. Finally, closed-form solution for the special case when $\rho = 1$ is reported in the main body of the paper.■

Proof of Proposition ACM (U) First, notice that for any pairs j, j' and e, e', equation (5)

$$\left[\frac{p_t(e,j)/p_t(e,j')}{p_t(e',j)/p_t(e',j')}\right] = \left[\frac{\left[A_t(e)\,C_t(e,j)\right]/\left[A_t(e)\,C_t(e,j')\right]}{\left[A_t(e')\,C_t(e',j)\right]/\left[A_t(e')\,C_t(e',j')\right]}\right]^{\theta}.$$

The pure absolute terms $A_t(e)$ and $A_t(e')$ cancel out within the ratios in the numerator and denominator, respectively. Hence,

$$\left[\frac{p_t(e,j)/p_t(e,j')}{p_t(e',j)/p_t(e',j')}\right] = \left[\frac{C_t(e,j)/C_t(e,j')}{C_t(e',j)/C_t(e',j')}\right]^{\theta}. \tag{18}$$

Then, for any $\bar{C}_t > 0$, a constant across e and j, we have $C_t(e,j) = \bar{C}_t * p_t(e,j)^{\frac{1}{\theta}}$ solves the solution and, without loss of generality, we can normalize $\bar{C}_t = 1$, so all absolute advantage terms are scaled in $A_t(e)$. To see that the only solutions are given by this form, assume that there is another solution of the form $C_t(e,j) = \nu_t(e,j) * p_t(e,j)^{\frac{1}{\theta}}$ for some $\nu_t(e,j) > 0$ that varies across e and/or j. Then, using (18), it has to be the case that, for all e and j

$$1 = \left[\frac{\nu_t(e,j)/\nu_t(e,j')}{\nu_t(e',j)/\nu_t(e',j')}\right]^{\theta},$$

which can only hold if $\nu_t(e,j) = \nu_t^E(e)\,\nu_t^J(j)$ for some $\nu_t^E(e) > 0$ and $\nu_t^J(j) > 0$. Hence, without loss of generality, we can normalize $A_t(e)$ and $M_t(j)$ as respectively embedding those terms. This establishes part (a).

To establish part (b), plug $T_t(e,j) = A_t(e)\,p_t(e,j)^{\frac{1}{\theta}}$ into $p_t(e,j) = \frac{[w_t(j)T_t(e,j)]^{\theta}}{\sum_{i=1}^{J}[w_t(i)T_t(i,e)]^{\theta}}$ to obtain that for all e and j:

$$p_t(e,j) = \frac{\left[w_t(j)\,A_t(e)\right]^{\theta}p_t(e,j)}{\sum_{i=1}^{J}\left[w_t(i)\,A_t(e)\,p_t(e,i)^{\frac{1}{\theta}}\right]^{\theta}}.$$

Notice that the terms $A_t(e)$ cancel out, as absolute advantages do not drive the allocation of workers across occupations. More interestingly, $p_t(e,j)$ cancels across both sides of this equation and then, the conditions boil down to

$$1 = \frac{[w_t(j)]^{\theta}}{\sum_{i=1}^{J}[w_t(i)]^{\theta} p_t(e,i)}. \tag{19}$$

The only solution for these equations, as long as $p_t(e,i) \neq p_t(e',i)$ for some e, e' and i, is that

$$w_t(j) = w_t(i), \tag{20}$$

for all i and j. To solve for the vector of unitary wages $[w_t(1), ..., w_t(J)]$, plug $T_t(e,j) = A_t(e) p_t(e,j)^{\frac{1}{\theta}}$ into the expression $H_t(j,e) = S_t(e) p_t(e,j) \left\{\Gamma\left(1-\theta^{-1}\right) T_t(e,j) p_t(e,j)^{-1/\theta}\right\}$, and then sum over e to obtain

$$H_t(j) = \Gamma\left(1-\theta^{-1}\right) \sum_{e=1}^{E} S_t(e) A_t(e) p_t(e,j). \tag{21}$$

Since $w_t(j) = \bar{w}_t \times M_t(j) [H_t(j)]^{\rho-1}$, then, from (20)

$$M_t(j) \left[\sum_{e=1}^{E} S_t(e) A_t(e) p_t(e,j)\right]^{\rho-1} = M_t(i) \left[\sum_{e=1}^{E} S_t(e) A_t(e) p_t(e,i)\right]^{\rho-1},$$

or

$$M_t(i) = M_t(j) \frac{\left[\sum_{e=1}^{E} S_t(e) A_t(e) p_t(e,i)\right]^{1-\rho}}{\left[\sum_{e=1}^{E} S_t(e) A_t(e) p_t(e,j)\right]^{1-\rho}},$$

But, since $M_t(j)$ are distributional shifts in the CES for H_t, it has to be the case that $\sum_{i=1}^{J} M_t(i) = 1$. Writing all $M_t(i)$ in terms of a single $M_t(j)$

$$\sum_{i=1}^{J} M_t(j) \frac{\left[\sum_{e=1}^{E} S_t(e) A_t(e) p_t(e,i)\right]^{1-\rho}}{\left[\sum_{e=1}^{E} S_t(e) A_t(e) p_t(e,j)\right]^{1-\rho}} = 1,$$

and taking j terms out of the summation and solving, we get

$$M_t(j) = \frac{\left[\sum_{e=1}^{E} S_t(e) A_t(e) p_t(e,j)\right]^{1-\rho}}{\sum_{i=1}^{J}\left[\sum_{e=1}^{E} S_t(e) A_t(e) p_t(e,i)\right]^{1-\rho}}, \tag{22}$$

as claimed.■

Proof of Proposition H (U) With the expression for H_t

$$H_t = \left[\sum_{j=1}^{J} M_t(j) [H_t(j)]^{\rho}\right]^{\frac{1}{\rho}},$$

and plugging the expression for (22) and (21)

$$H_t = \left\{\sum_{j=1}^{J} \frac{\left[\sum_{e=1}^{E} S_t(e) A_t(e) p_t(e,j)\right]^{1-\rho}}{\sum_{i=1}^{J}\left[\sum_{e=1}^{E} S_t(e) A_t(e) p_t(e,i)\right]^{1-\rho}} \left[\Gamma\left(1-\theta^{-1}\right) \sum_{e=1}^{E} S_t(e) A_t(e) p_t(e,j)\right]^{\rho}\right\}^{\frac{1}{\rho}},$$

which, after re-arranging, grouping and simplifying

$$H_t = \Gamma\left(1-\theta^{-1}\right) \left\{\frac{\sum_{j=1}^{J}\sum_{e=1}^{E} S_t(e) A_t(e) p_t(e,j)}{\sum_{i=1}^{J}\left[\sum_{e=1}^{E} S_t(e) A_t(e) p_t(e,i)\right]^{1-\rho}}\right\}^{\frac{1}{\rho}},$$

as claimed.■

Data

In this appendix, we described in more detail the data we used for our calculations. All data is extracted from the *Integrated Public Use Microdata Series, International (IPUMS).* IPUMS consists of microdata samples from international census records. The records are converted into a consistent format and made available to researchers through a web-based data dissemination system by the Minnesota Population Center. Based on data availability, we choose 11 specific country-year combinations: Brazil 2000; India 1999; Mexico 1960 and 1990; United States 1960 and 2000; Indonesia 1976 and 1995; Panama 2000; Canada 1971 and 2001.
From IPUMs, we collect the following variables:

- `age`: Respondent´s age in years
- `sex`: Respondent´s gender
- `edattain`: Respondent´s educational attainment. Specifically, the highest level of schooling completed (degree or other milestone). The emphasis on completion is critical: a person enrolled in the final year of secondary school only receives the code for having completed lower secondary only – and in some samples only primary. `edattain` is an attempt to merge – into a single, roughly comparable variable – census sources that provide degrees, others that provide actual years of schooling, and those that have some of both. .
- `yrschool`: Respondent's completed years of schooling, regardless of schooling type. Only formal schooling is counted. Top coding is frequent.[15]
- `empstat`: Indicates whether or not the respondent was part of the labor force – working or seeking work – over a specified period of time prior to the census. This variable classifies all respondents into three groups: employed, unemployed, and inactive. The combination of employed and unemployed yields the total labor force.
- `occisco`: Respondent's primary occupation, coded according to the major categories in the 1988 International Standard Classification of Occupations (ISCO) scheme. For someone with more than one job, the primary occupation is typically the one in which the person had spent the most time or earned the most money. Occupation categories are:
  - Legislators, senior officials and managers
  - Professionals
  - Technicians and associate professionals
  - Clerks
  - Service workers and shop and market sales
  - Skilled agricultural and fishery workers
  - Crafts and related trades workers
  - Plant and machine operators and assemblers
  - Elementary occupations
  - Armed forces

[15] Top codes are figures that do not represent the real variable value, but instead are set because there is an upper boundary in the census data. For example, with `yrschool`, in Indonesia 1976 the variable does not take a value higher than 16, even though a person may have more than 16 years of formal education.

More information on occupations can be obtained at http://www.ilo.org/public/english/bureau/stat/isco/isco88/index.htm

- `incearn`: Respondent's total labor income in the previous month or year. Does not only include wages, but also income from businesses and farms. Most samples report data for the previous month. Data for Canada and the United States are annual figures. Amounts are expressed as they were reported at the time of the census in the currency of the respective country. Some samples report negative earnings for individuals. Top coding is frequent.[16]

- `incwage`: Reports the respondent's weekly, monthly or annual wage and salary income. The data are recorded in the currency of each country in that census year and are not adjusted for inflation. Topcodes are frequent, and vary across countries[17]. The data for the United States and Canada are annual figures; the data for Indonesia and Panama are monthly; other samples report weekly figures.

.1 Adjustments on the Raw Data and Other Calculations

Following the model, we adjust the data to obtain a sample that fits the model definitions, and that allows for comparisons across time and countries. This implies using the data to get measures of occupation, income, education attainment, gender, and age for all individuals.

To characterize occupations, we drop all individuals that do not fit in one the nine occupation categories: Managers, Professionals, Technicians, Clerks, Services, Agriculture, Traders, Operators and Elementary. These includes individuals with other occupations, individuals with unknown occupations, and individuals where the category is "not in universe" (NIU), such as children. We also exclude respondents occupied in the Armed Forces, as the market mechanisms that underlies the allocation of workers across occupation might not fully apply to the military, particularly in the presence of conscription.

For income, we use the variables `incearn` and `incwage` to measure individual wages for each country. Unfortunately, for some samples `incearn` was present but not `incwage,` and vice-versa. Therefore, the exact variable for each sample depends on availability, `incwage` being preferred when that both are available. Individuals with non-positive values, missing values, or top codes are eliminated from the database. The elimination of top-coded individuals introduces an inevitable bias in the data, as we exclude top earners from the sample. However, we think that is better than the alternative, as we would otherwise measure the income of top-coded individual with significant error, and introduce artificial equality across all top-coded respondents in a particular sample. In the particular case of Mexico 1960, a significant number of individuals had `incearn` = 1. In regards to the issue, IPUMS states that "*A large number of cases in 1960 have a value of "1." It may indicate a low income value, but almost certainly does not literally mean 1 peso."* Following the same logic as with top-coded earnings, these observations are eliminated, as we do not know the true value of these earnings. Therefore incomes in Mexico 1960 may be biased upwards.

Additionally, we transform income variables to annual equivalent year 2000 USD PPP figures. First, as incomes are reported in different frequencies, we multiply monthly (weekly) incomes by 12 (52). Second, we use CPI inflation[18] for each country to adjust nominal incomes in a given year into equivalent incomes for the year 2000. Finally, we convert equivalent incomes in local currency to equivalent USD adjusted by GDP PPP values using the 2000 PPP conversion factor[19] (also known as the PPP exchange rate)..

Education measures are derived from `edattain and yrschool.`From here we define seven education categories: (1) no schooling, (2) incomplete primary, (3) complete primary, (4) incomplete secondary, (5) complete secondary,

[16]In this case, top codes represent, where applicable, a determination by the Current Population Survey that some high values were too sparse and specific to be recorded as they were reported to the CPS without the possibility of identifying the respondents. In these cases, the CPS put numerous high value cases together under one particular high value to protect respondent anonymity.

[17]Same as `incearn`

[18]Data from the World Bank: http://data.worldbank.org/indicator/FP.CPI.TOTL.ZG

[19]Data from the World Bank: http://data.worldbank.org/indicator/PA.NUS.PPP?view=chart

(6) incomplete tertiary, and (7) complete tertiary. Individuals with missing observations are dropped. Minor adjustments must be done for some countries in which the denomination of education categories differs slightly.

For demographics, we define three age groups: "young" individuals aged between 25 and 35 years old, "middle aged" individuals between 36 and 50, and "old" individuals with ages above 50. Individuals below 25 or with missing data on age are dropped from the samples. Individuals with missing data on gender are also dropped.

Finally, as the model does not consider leisure, home production, or unemplomet, we only keep observations for employed individuals, given by `empstat`. This filter was not applied for Mexico 1960, as that sample has no information for `empstat`. This may bias the data for Mexico 1960 by including unemployed or inactive individuals that still report positive income and the other variables.

With the final data for each sample, all combinations of gender (two groups), age (three groups) and education attainment (seven groups) are used to create 42 human capital groups e. Those 42 groups are combined with the 9 occupation categories to construct 378 group-occupation combinations for each sample. In each sample, group-occupation combinations that have less than ten individuals are eliminated to avoid outliers which might bias the results.