A Division of Laborers: Identity and Efficiency in India

Guilhem Cassan
University of Namur

Daniel Keniston
Louisiana State University

Tatjana Kleineberg
World Bank[*]

December 29, 2020

Abstract

Workers' social identity affects their choice of occupation, and therefore the structure and prosperity of the aggregate economy. We study this phenomenon in a setting where work and identity are particularly intertwined: the Indian caste system. Using a new dataset that combines information on caste, occupation, wages, and historical evidence of subcastes' traditional occupations, we show that caste members are still greatly overrepresented in their traditional occupations. To quantify the effects of caste-level distortions on aggregate and distributional outcomes, we develop a general equilibrium Roy model of occupational choice. We structurally estimate the model and evaluate counterfactuals in which we remove castes' ties to their traditional occupations: both through their direct preferences, and also via their parental occupations and social networks. We find that the share of workers employed in their traditional occupation decreases substantially. However, effects on aggregate output and productivity are very small–and in some counterfactuals even negative–because gains from a more efficient selection based on individuals' comparative advantage are offset by productivity losses due to weaker caste networks and reduced learning across generations. Our findings emphasize the importance of caste identity in coordinating workers into occupational networks, thus enabling productivity spillovers.

[*] The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent. This work was supported by the Fonds Wetenschappelijk Onderzoek – Vlaanderen (FWO) and the Fonds de la Recherche Scientifique – FNRS under EOS EOS project O020918F (EOS ID 30784531). We are grateful to Joseph Altonji, Mark Rosenzweig and Nicholas Ryan and audiences at NES, WEHC, LSU, UBC, UNamur, ULB, BREAD, Warwick, Chicago, USC, Tulane and SEA for helpful suggestions. We are extremely grateful to James Nye at the University of Chicago library for his invaluable help in accessing the 1911 Census data.

"...the Caste System is not merely a division of labour. It is also a division of labourers."

"... the Caste System [...] involves an attempt to appoint tasks to individuals in advance, selected not on the basis of trained original capacities, but on that of the social status of the parents."

Ambedkar (1936)

1 Introduction

Work is more than a source of income: it is a part of identity and subject to social norms. Thus occupational choices are not purely economic, but rather the outcome of an individual's ethnic background, personality, and social aspirations. While the complexity of the occupational choice problem is widely recognized, the challenges of quantifying its non-economic factors have presented substantial barriers to estimating their importance. In this paper, we analyze occupational choices in the context of the Indian caste system. Each Indian caste is associated with (usually) a single traditional occupation, which was historically seen as the proper vocation for members of that caste in society–their "dharma. While the end of the link between occupation and caste has often been predicted (Srinivas, 2003), caste remains salient for individuals in modern India. Traditional occupations, which are exogenous from the perspective of any single individual, provide a unique opportunity to study the role of identity in occupational choice in a context where an important element of identity is observable and predetermined.

The goal of this study is to quantify both the importance of identity for an individual's occupational choice, as well as the impact of these identity-influenced choices on the economy as a whole. Caste-based occupation identity affects the aggregate economy via two distinct channels. First, the preference for one's traditional occupation can distort individuals' selection into occupations away from their comparative advantage, leading to an inefficient allocation of talent across occupations. Second, the sorting of castes into traditional occupations shapes the economy in ways that affect workers' choices well beyond their own preferences. The key channels of this "path dependence" mechanism are human capital inherited from parents (many of whom were employed in traditional occupations) and social networks (which are formed around traditional occupations). Such effects reinforce persistence in castes' occupational choices: even if workers no longer feel tied to their traditional occupations, they might nevertheless work in them to take advantage of the productivity effects of large caste networks and working in the same occupation as their fathers. Unlike the distortionary effects of occupational preferences on human capital allocation, the aggregate impact of these channels of historical persistence may be positive, and are essential to explaining the remarkable endurance of occupational identity over time.

We first document a series of new empirical facts that illustrate the role of caste membership for occupational choices. We find that individuals are about three times as likely to work in their traditional occupation compared to any other occupation. Within a caste, those workers employed in their traditional occupation earn less than their caste-mates who have given up their traditional occupation to work in other occupations. However, when we examine earnings within occupation (that is, controlling for occupation-level fixed effects) we find that workers employed in their traditional occupation earn more than workers from other castes who work in the same occupation. The data further shows that returns to ability–measured by schooling and experience– are lower in common traditional caste occupations compared to "modern" occupations. These empirical findings inform our model specification.

We then develop and estimate a structural general equilibrium Roy (1951) model of education and occupational choice that incorporates caste identity through several channels. The selection based on individuals' comparative advantage lies at the heart of this analytical approach. In a common Roy model specification, workers differ only on individual productivity which varies independently across occupations. In this framework, preference for traditional occupation pushes additional workers into occupations where they are less productive than workers who selected the occupation based solely on comparative advantage. We extend the model to allow workers to also differ in general ability, in which case utility from working in one's traditional occupation might induce the opposite type of selection. High general ability workers, who would otherwise prefer to work in "modern" occupations where returns are higher, may be drawn back into their traditional occupation. This extended Roy model, allowing workers to differ in both occupation-specific and general ability, can reconcile the negative within-caste and positive within-occupation wage effects of traditional occupation found in the data.

To study the importance of these channels and to quantify their aggregate effects, we consider the economy in a general equilibrium context. Since wages reflect the marginal product of human capital in an occupation, we cannot accurately estimate the effect of reallocation of human capital across occupations unless we allow wages to adjust. In addition, occupational choices are closely linked to individuals' education choices and to the composition of social networks. We therefore determine wages, education and social networks endogenously. In particular, the general equilibrium nature of our analysis allows for the possibility that castes' occupational identity can serve as a means of equilibrium selection, helping workers to coordinate on a human capital allocation and a network composition that maximizes output (Chen and Chen, 2011).

Identity is rarely expressed solely in terms of occupational preferences, and this is particularly true in the Indian system. The caste system is (perhaps primarily) a hierarchical structure (Ambedkar, 1936; Dumont, 1970), with certain groups traditionally seen as ritually superior and others historically viewed as "polluting". We control for the hierarchical ranking to avoid conflating the role of identity preferences and discrimination which may be correlated with the characteristics of traditional occupations. In addition, we allow women and workers from hierarchically lower castes to be subject to wage discrimination and heterogeneous costs of acquiring education. These model ingredients affect individuals' education and occupational choices, which in turn determine the stock and allocation of human capital, the wage rate per human capital unit in each occupation, and ultimately the output of the economy (via an aggregate production function).

To estimate the model, we construct a novel dataset that includes micro-data on occupational choices, wages, and demographics, and link this to detailed data on castes' traditional occupation from historical sources. We use our estimated model to investigate the importance of occupational identity on occupational distributions, wages, aggregate output, and aggregate productivity. We do so by successively removing channels that link castes' to their traditional occupations: we remove agents' attachment to their traditional occupations–first, holding caste networks constant, then endogenizing it as a function of workers' occupational choices.

In both cases, we find very small aggregate effects: output per worker increases by 0.6 percent and aggregate output by 0.3 percent. These effects are small because gains from workers' improved selection based on their comparative advantage is offset by reduced productivity from weaker caste networks and less intergenerational knowledge transfer (i.e. less workers work in their fathers' occupation). Despite the trivial aggregate effects, we nevertheless find large distributional effects: The share of traditional workers decreases by 4 to 5 percentage points in the most affected occupation and by roughly 6 percentage points in the most affected caste. Overall, the aggregate share of workers who work in their traditional occupation decreases by roughly 10 percent.

We then additionally remove the last link of castes to their traditional occupations by removing the correlation between fathers' occupations and traditional occupations. With constant caste networks, this leads to a large drop in output (-2.8 percent) but small gains in output per worker (0.08 percent). Output drops because we remove the complementarity between jointly choosing one's traditional occupation–with largest networks–and one's father's occupation. Again, these negative effects weigh against better selection of workers based on their comparative advantage. With endogenous networks, losses are even larger: total output drops by 8 percent and output per worker by 5 percent. Caste networks are now weaker since we have removed all coordination elements that previously clustered caste networks in their traditional occupations–lowering productivity and output. These findings show that reduced productivity through weaker networks and less parental learning can dominate over gains from better selection of workers based on their comparative advantage.

The paper proceeds as follows. Section 2 reviews the relevant literature. Section 3 and 4 describe the data and our reduced form analysis. Section 5 presents our model. Section 6 describes the estimation strategy. Section 7 presents the results and Section 8 concludes.

2 Literature Review

This paper contributes to several branches of the literature studying occupational choice and human capital allocation both in India and more generally.

Since at least Akerlof and Kranton (2000), a growing literature studies the role of identity in a range of economic decisions. Examples include recent studies on religious identity and contributions to public goods (Benjamin et al., 2016), religious identity and food consumption (Atkin et al., ming), criminal identity and cheating (Cohn et al., 2015), attachement to hometown and reduced occupational and geographical mobility (Munshi and Wilson, 2011), as well as national identity and support for redistribution (Shayo, 2009). Another relevant literature examines the intergenerational transmission of education and occupation choices. (Altonji and Dunn, 2000; Phelan and Kinsella, 2009) find for example that parents and communities play important roles in shaping young adults' occupational choices. While the concept of "occupational identity" is central in Akerlof and Kranton (2000) and has attracted a large literature in Western sociology and psychology (see Skorikov and Vondracek (2011) for a recent review), it received relatively less attention from economists, in particular in empirical work. Several theoretical works, however, have looked into the role of social norms in economic behavior. Akerlof (1980) for example shows that the fear of loss of reputation may prevent individuals from making economically optimal choices on the labor market if it entails the disobedience of social norms. Anticipating our results, Akerlof (1976), shows that in the presence of widely shared social norms, even the removal of a taste for following these norms may not be enough to alter behaviour due to the fear of social sanctions.

Our paper directly adds to the rich literature on the interaction between the Indian caste system and occupational identity. Sociologists have suggested that occupational links are the origin and defining feature of the Indian caste system at the *jati* level (Gupta, 2000)[1], which is the most relevant dimension of caste identity for most Indians (Vaid, 2014). Oh (2019) studies this link formally using an experimental methodology, offering workers casual labor tasks that are either linked to their traditional occupation or associated with a different caste. She finds workers are significantly less likely to accept casual labor outside of their traditional occupation, especially if the task is associated with a hierarchically inferior caste.

Economic studies have further emphasized the importance of castes as occupational networks. In particular, Munshi and Rosenzweig (2006) examine the educational choices of Mumbai residents, arguing that lower castes discourage their most able members from pursuing high skill occupations in order to preserve strong social networks in low skill traditional occupations. An excellent survey of the economics literature on caste is provided by Munshi (2019).[2] Our paper complements this literature by formally analyzing the implications of the Indian caste system for human capital allocation and aggregate output in a general equilibrium model where occupational wages and networks adjust endogenously.

In the Indian context, several studies have documented that intergenerational transmission of education (Borkotoky et al., 2015) and occupation (Kumar et al., 2002; Deshpande and Palshikar, 2008; Vaid, 2012; Hnatkovska et al., 2013; Iversen et al., 2017) is remarkably strong. Kumar et al. (2002) and Vaid (2012) have also documented that the role played by caste on intergenerational transmission of occupation did not seem to weaken much over time.

Our work also relates to recent studies that explore the aggregate implications of frictions to human capital allocation. Perhaps the work most similar in spirit is the paper by Hsieh, Hurst,

[1] See also the literature on the pattern of caste based patron-client relationships and occupational specialization, the "jajmani" system, iniated byWiser (1936).

[2] The long-term case study of the Palanpur Village (Lanjouw and Stern, 1998) contains a detailed description of the caste roles in the village. Researchers in Palanpur anticipate the results of our broader analysis in their demonstration that traditionally agricultural castes perform better in agriculture than others, and are less likely to enter off-farm occupations.

Jones, and Klenow (2013), which quantifies the effect of decreased discrimination against women and blacks in high skill-return US occupations on aggregate wages and GDP growth. It is common in this literature to assume that occupation-specific talent is uncorrelated across occupations. Under this assumption, workers' selection is such that expanding sectors increasingly attract individuals with lower comparative and absolute advantage, while contracting sectors shed the least productive workers (as noted by Young in his 2014 study of US industries). Lagakos and Waugh (2013) use this insight to study the cross country relationship between output and agricultural productivity. Alvarez-Cuadrado et al. (2019) re-examine the selection into agriculture at the micro level, and find that individuals who are more productive at farming are also more productive in their secondary occupations.[3] We add to this literature by empirically testing the implications of different model assumptions on occupational choices and the wage distribution across and within castes.

3 Data

To study the effects of identity on human capital allocation and aggregate output, we need a dataset that contains detailed information on workers' identity characteristics, their occupational choices and wages. In the Indian context, workers' identity and social network is defined by their sub-caste, or *jati,* rather than by the larger *varna* caste groupings or the government reservation categories that group different castes. Datasets that contain information of workers' jati have only recently become available. The primary dataset used in this project is the Indian Household Development Survey (IHDS), which provides detailed information on individuals' demographics, occupations, wages, and family characteristics. We complement this dataset with two crucial additional data moments. First, we construct social networks at the jati-occupation level from the Demographic and Health Survey (DHS) (IIPS, 2007). Second, we retrieve information on each jati's traditional occupation from the colonial Census in 1911–which we again complement with other historical sources. Merging these three datasets at the jati level poses particular challenges and required a very labor-intensive harmonization of jati names. In the following paragraphs, we first explain our strategy of merging jati names across datasets before briefly describing the three main data sources in more detail.

Harmonization of Jati names:

Both the IHDS and DHS, report jati names declared by respondents verbatim. This complicates classification, first because the meaning of "caste" itself can be ambiguous (Headley, 2013), and second because there are many synonyms and spellings for each jati, not to mention typos. To clean and harmonize jati names in a systematic manner, we use the People of India project–launched in 1985 by the Anthropological Survey of India, which has been an extraordinary effort to systematically collect data on all jatis of India. This project produced a volume (Singh, 1996), listing

[3] In spatial general equilibrium, Eckert and Peters (2018), Heise and Porzio (2019) and (Bryan and Morten, 2018) find that the "uncorrelated shock" Roy model can not explain the patterns of regional migration and productivity in the data.

all jati names and their various synonyms at the state level. We digitized this volume to create a jati "master list" with state-specific lists of all jati synonyms. We then hand-merged this master list with the IHDS and DHS with the help of several research assistants, ultimately categorizing 32,137 recorded names into 2,650 unique castes.

Individual-level data from IHDS Household Survey:

The primary dataset used in this project is the second round of the Indian Household Development Survey (IHDS), which was conducted in 2011 (Desai et al., 2008; Desai and Vanneman, 2015). This dataset contains rich demographic data on 42,152 households, including extensive occupation and income module that record income and time spent in each occupation for each individual of the household.[4] The survey also documents the occupation of household head's father.[5] When missing, we impute father's occupational probabilities based on the data of fathers' occupation for individuals in the same caste.[6] We add information on jatis' social status by matching the jati names in our dataset to their contemporary classification as Scheduled Castes (SC), Scheduled Tribes (ST) or Other Backward Classes (OBC), which are rough proxies of social ranking. Following Cassan (2019), we add information on jatis' social status by matching the jati names in our dataset to their contemporary classification as Scheduled Castes (SC), Scheduled Tribes (ST) or Other Backward Classes (OBC), according to official state-wise lists of reservation classifications.

Occupation-specific caste networks from DHS Household Survey:

While the IHDS data is extensive, the data requirements to estimate occupation-level social networks for over 1,000 jatis and 48 occupations are very demanding on the sample size. We therefore use the third round of the DHS (2005-06) (also called NFHS), to provide caste and occupation information in a much larger additional sample of 109,041 households. However, because the DHS contains only very sparse information on income and parental occupation, we use it only to supplement the social network data while relying on the IHDS data for the main part of our analysis.

Traditional occupations from historical data sources:

Our main source of information on jati's traditional occupations is the colonial Census of 1911, which lists the traditional occupation of each jati in each province (Conlon, 1981). We complement the data with several other historical data sources to improve the completeness of the dataset.[7] To create

[4] The IHDS is a panel dataset with two rounds, however, we use the first round (2005-06) only to complement missing or incomplete data from the second round, when necessary. In particular, we use the jati name from the first round if the jati name in the second round is missing or coded in a very general way such as "scheduled caste" (SC). Similarly, we use parental occupation from the first round if it is missing in the second round. The income and time use data on secondary occupations, home work, animal care, money lending, and land rental posed specific challenges in the cleaning and construction of our final dataset, which we explain in detail in Appendix A1.

[5] If the head of the household is a women, the survey records the occupation of her husband's father.

[6] Overall, information on father's occupation is missing for 12.8 percent of men and for 84.7 percent of women.

[7] Our primary source are the 1911 Tables titled, "Occupation by selected castes, tribes or races". If jatis are missing in the 1911 Census, we use data from Kitts (1885), which is based on the 1881 Census. If jatis are missing in both

a crosswalk between historical occupation classifications and their contemporaneous counterparts used in the IHDS and DHS datasets, we create 48 consistent occupational categories (see list in Appendix Table A2).

4 Reduced Form Evidence

Traditional occupation and occupational choice

We start by measuring the extent to which the traditional occupation of a jati determines the contemporary occupational choice of its members. Figure 1a documents the share of male workers in each occupation who work in their jati's traditional occupation and the share who works in their father's occupation. Figure 1b does the same for females. Both show large heterogeneity across occupations. Some occupations are predominantly done by workers who follow their jati's traditional occupation–such as for example "dyeing and cleaning" for which half of all workers follow their traditional occupation. Other occupations have close to no traditional workers–such as legal or medical professions. Workers also have a tendency to follow their father's occupation, however, this is much less marked for women. Occupations with a high share of traditional workers also tend to have more workers who follow their father's occupation, however, these mechanisms are not perfectly correlated. Overall, 16.8% of men work in their jati's traditional occupation, versus 9.1% if workers were randomly allocated across occupations (keeping the occupational distribution constant).[8] For women, 8% are in their jati's traditional occupation against 5% if randomly allocated.[9]

These findings support the hypothesis that caste identity is closely linked to traditional occupations. An alternative explanation could be that consumers prefer products which are provided by members of the traditional caste (i.e. from castes whose vocation it is to produce that good). This may be true in some cases,[10] however, Figures 1a and 1b show that many occupations in which the identity of the producer is unknown to consumers remain strongly associated with their traditional castes–for example, fishing, jewelry, or cultivation. Oh (2019) also finds in an experimental study that workers' preference for their traditional occupation is not affected by whether occupational choices are made in public or private.

To quantify the effect of traditional occupational preferences more formally, we turn to a regression analysis. We rectangularize our dataset at the individual×occupation level, so that each individual is observed 48 times (once for each potential occupation). The variable of interest

datasets, we use the People of India "India's Communities" volume which provides rich historical and anthropological information about all jatis–usually including jatis' traditional occupation. For jatis whose traditional occupation was labeled as "criminal" in colonial era sources, we found historical evidence that these groups were nomadic and subject to state-level discrimination that aimed at sedentarizing those groups (Schwarz, 2010). For these jatis, we instead retrieve their traditional occupation from data in Crooke (1896) or–if missing there–from the People of India volume.

[8] The effects are larger for men in rural areas with the respective numbers being 19.9% and 13% for rural ares and 13% and 4% for urban areas.

[9] Again, the effects are larger in rural areas with the respective numbers being 11.4% and 8% for rural women and 2.1% and 0.8% for urban women.

[10] For example in the case of Hindu religious workers of the priestly caste. However, only 93 workers out of our sample of 98,344 individuals are employed as religious workers.

Occ_{iok}, indicates that individual i of jati k works in occupation o. We then run the following OLS regression:

$$Occ_{iok} = \alpha + \beta TradOcc_{ok} + \gamma X_o + \delta Z_i + \varepsilon_{iok},$$

where $TradOcc_{ok}$ indicates that occupation o is jati k's traditional occupation, and X_o and Z_i are occupation and individual fixed effects. We cluster standard errors at the PSU level in accordance with the 2-stage sampling of the IHDS, following Abadie et al. (2017). Table 1 presents the results. Column 1 of Panel A shows that men are 6.8 percentage points more likely to work in an occupation if that occupation is their jati's traditional occupation, holding constant occupational and individual characteristics. In column 2 we include individuals' jati network, defined as the share of workers in their chosen occupation that belongs to their caste, and an indicator for whether they work in the same occupation as their father as additional covariates. We find that both variables are significant and large in magnitude: workers are roughly 31 percentage points more likely to choose their father's occupation; and 10 percentage points more likely for each 1 percentage point increase in their caste-occupation network. The impact of traditional occupations remains significant but is lowered to roughly 4 percentage points, which implies that workers are three times more likely to work in their traditional occupation compared to a random occupational choice. Male scheduled castes (SCs) have less affinity for their traditional occupations, although these results are imprecisely estimated (see column 4).

In Panel B of Table 1 shows that these effects are present but much smaller for women. On average, women are 2.7 percentage points more likely to work in their traditional occupation, which reduces to 0.7 percentage points when controlling for intergenerational transmission of occupation and networks. Women's occupational choice probability increases by roughly 14 percentage points for their father's occupation and by roughly 3 percentage points for each 1 percentage point increase in the occupation's caste network.

Selection and productivity in traditional occupations

We examine the relationship between occupational identity and income. We run the following regressions:

$$\log(wage/hour)_{iok} = \alpha + \beta TradOcc_{iok} + \gamma X_{iok} + \varepsilon_{iok}$$

where $\log(wage/hour)_{iok}$ is the log of hourly wages of individual i from jati k working in occupation o, $TradOcc_{iok}$ is a dummy indicating whether individual i's occupation o is her jati k's traditional occupation, and X_{iok} a set of individual characteristics. In Table 2 we present the results and we consider two specifications, controlling first for jati fixed effects and then for occupation fixed effects. We control for father's occupation and caste networks in all regressions.

With jati fixed effects (columns 1 and 3), we find that male workers earn lower hourly wages in their traditional occupation, compared to workers from the same jati who work in any other (non-traditional) occupation. This finding is consistent with the standard selection effects of the Roy model: because workers get a utility boost, they are willing to accept lower wages in their traditional occupation.[11] The result also holds for women with a slightly smaller coefficient (see column 3), conditioning on labor force participation.[12]

Perhaps surprisingly, with occupation fixed effects (columns 2 and 4), the results are reversed. Here we find that workers in their traditional occupation earn 12-13 percent more per hour than non-traditional workers in the same occupation. This result is at odds with the logic of the standard Roy model: if we assume uncorrelated occupational skills, then traditional workers should have lower average productivity and hence lower hourly wages than other workers in the same occupation. Additionally, we find stronger within-occupation effects of parental occupation and social networks for women (column 4) than men (column 2), again conditioning on labor force participation.

Returns to ability in traditional occupations

To explain these findings, we examine differences in returns to ability as an potential driver of occupational choices. We conjecture that traditional occupations–which existed by definition in pre-industrial times–have different returns to ability than "modern" occupations. We examine the conjecture of differential returns to ability in traditional occupations with the following regression:

$$\log(wage/hour)_{iok} = \alpha + \beta TradOcc_{io} + \gamma X_{iok} + \delta TradOcc_{io} * X_{iok} + \varepsilon_{iok},$$

where X_{iok} are years of schooling and experience, measuring workers' general ability. Note that we now define $TradOcc_{io}$ as traditional occupations of *any* jati (hence, it is not indexed by caste k) to examine the characteristics of traditional occupations at the occupation level. We are particularly interested in coefficient δ which captures differential returns to schooling and experience in traditional occupations. All regressions include caste and occupation fixed effects, and are estimated separately by gender and conditional on labor force participation.

Table 3 column 1 shows that returns to ability are indeed lower in traditional occupations for male workers: wages increase by 7.4 percent per year of schooling in non-traditional occupation and by only 4.2 percent in traditional occupations. Similarly, returns to experience are more than twice as large in non-traditional occupations. In column 2, we examine effects from father's occupation and caste networks. Both variables increase wages significantly and substantially in magnitude, but we find that returns are not significantly lower in traditional occupations. Including all variables together in column 3 does not change the previous findings.

For women (columns 4-6), we find much lower and imprecisely measured returns to education and experience in all occupations. For returns to fathers' occupation and caste networks, we again

[11] These results provide evidence against the hypothesis that consumers, or intermediaries, have a higher willingness to pay for products sold by traditional workers. If that were the case, workers would earn more in their traditional occupation than their fellow caste members in other occupations.

[12] We omit home workers from the regression as they have no income data. This primarily affects the sample size for women.

find larger effects for women and general and no significant difference in traditional occupations.[13]

Discrimination

Last, we examine whether there is wage discrimination by gender or by castes' social ranking. We proxy the latter by castes' categorization into Other Backward Classes (OBC), Scheduled Castes (SC), Scheduled Tribes (ST).[14] All specifications control for individual characteristics, including education and experience. Table 4, column 1 shows very large discrimination effects on women (-0.59 log-points) and much smaller effects for castes lower in the traditional hierarchy. In column 2, we add occupation fixed effects. Within occupations, we find smaller estimates for women (albeit still large at -0.38 log-points) but we find larger and more precisely estimated discrimination effects against lower-ranked castes (-0.07 log-points for OBCs, -0.18 for SCs and -0.24 for STs). The estimates are robust to controlling for father's occupation, caste networks, and a traditional occupation dummy (columns 3 and 4).

Summary

Our reduced form results show that preferences for traditional occupations are important for occupational choices.

Yet the standard selection in a Roy model with uncorrelated occupational talent cannot fully explain the wage patterns that we observe for workers in their traditional occupation.[15] In the uncorrelated Roy model, as emphasized by Young (2014), the first workers to enter an occupation have the highest occupation-specific ability, so that the average occupation-specific ability of a caste would decrease if more of its workers enter an occupation due to traditional occupation preference. This finding is consistent with the empirical results in Columns 1 and 3 of Table 2: within a caste, those working in their traditional occupation earn less. If workers are drawn to an occupation due to higher productivity (rather than preference) this will offset the negative selection, and indeed we see no significant difference in wages for workers in their fathers occupations in columns 1 and 3.

However, when comparing workers within an occupation we find traditional workers earn more than outsiders–the opposite of the comparative static predicted by the uncorrelated Roy model. To understand this result, suppose that workers also differ in their general human capital, and that (as we show above) traditional occupations have relatively lowe returns to this general ability. High-

[13] All results are qualitatively the same, and in some cases more precise, if we restrict our definition of "traditional occupations" to occupations that are traditional for a minimum share of the population (e.g. for more than 0.5 percent).

[14] These groups are eligible for affirmative action policies by the Indian government. SCs were historically most discriminated, STs are aboriginal tribes with limited access to public goods, OBCs are low in the caste hierarchy but were subject to less discrimination than SCs.

[15] The reduced form analysis could be biased by the differential selection of castes into occupations with heterogeneous characteristics. We address a variety of potential concerns through a set of robustness tests. We control for other potential determinants of occupational choice, such as the traditional "purity" of an occupation, and its interaction with a caste's hierarchical status. We also examine the role of inherited land, and the occupational choices of other family members. While these variables are often significant, they cause virtually no change in the estimated effects of traditional occupation. See Section A2.1 for further discussion and A1 for results.

ability workers then sort a priori into modern occupations and low ability workers into traditional occupations. Those marginal workers who are drawn into their traditional occupation due to the utility boost have higher ability than the average (non-traditional) workers in the same occupation. This mechanism allows us to rationalize the empirical results from Columns 2 and 4 of Table 2: within traditional occupations, traditional workers earn more than non-traditional workers due to higher general ability. Intuitively, some high skilled individuals continue working in their low skill-returns traditional occupations (agriculture, laundering, pottery, etc.) when, in the absence of the caste-occupation affinity, they might apply their skills more productively in high-returns occupations (teaching, engineering, law, etc.). Guided by these empirical findings, we specify our general equilibrium occupational choice model to study the importance of caste identity on aggregate outcomes.

5 Model

We now first describe the model setup. We then solve agents' education and occupational choices. Last, we present the production side and market clearing.

5.1 Model Setup

Individual Characteristics:

Individuals i differ in general ability or talent, which consists of an unobservable component α_i and an observable component β_i. In addition, individuals receive an idiosyncratic education cost shock η_i and a vector of idiosyncratic occupation-specific productivity shocks π_{io}.

Caste Affiliation and Family Environment:

Each individual belongs to a caste k that affects their utility payoffs and choices via four channels. First, agents have a direct preference for working in their caste's traditional occupation, which we denote by τ_{ok}. Second, members of caste k can experience wage discrimination, which we denote byT_k. Third, we allow for productivity effects from social networks, which we define as the share of all workers in an occupation that are members of his caste k. Last, caste affiliation can affect costs of schooling, which we denote by κ_k. These costs are measured in utils per year and can capture the pecuniary cost of schooling (school fees, scholarships, etc.) as well as non-pecuniary factors such as potential caste-level discrimination, returns to education on the marriage market, or other social norms that make schooling more or less costly for certain castes.

In addition, we allow for productivity effects from working in the same occupation as one's father since parents can transfer skills, customer networks, or other assets to their children.[16] To simplify notation, we denote the total productivity shifter from both caste networks and father's occupation for an individual i in occupation o by ψ_{io}.

[16] We assume that these intergenerational effects only exist when children work in the same occupation as their father and that they are zero otherwise.

Occupation Characteristics:

Each occupation offers a wage rate w_o per human capital unit, which is endogenously determined. Occupations differ in their returns to general skill, which we denote by ρ_o and which models the inherent skill-intensity of an occupation (e.g. engineering is more complex and skill-intensive than agricultural labor). We further allow occupations to vary in their amenities A_o, which captures that some occupations might be more pleasant than others as well as any entry cost that is not directly measurable in the wages of an occupation. In the Indian context, examples of such entry costs might include exams to enter government service, or high costs of acquiring farm land due to imperfect land markets.

Preferences:

Workers have preferences for the homogeneous consumption good C, for working in their caste's traditional occupation τ_{ok}, and for the amenities of their occupation, A_o. We assume a log-linear functional form, so that the utility of a worker i from caste k who works in occupation o is equal to:

$$U_{io} = \log(C_{io}) + \tau_{ok} + A_o. \tag{1}$$

5.2 Education and Occupation Choices

Timing of choices: Individuals live two periods: childhood and adulthood. At birth, they know their caste affiliation, their general skill α_i and their education taste shock η_i. Individuals first choose years of schooling s_i, which remain fixed during adulthood and is a component of workers' general ability level β_i. Young adults then receive idiosyncratic occupation-specific productivity shocks π_{io} and choose an occupation o in which they work during adulthood. Occupation-specific shocks are realized only after education is completed, so children take expectations over these shocks when choosing education. We solve the problem backwards, beginning with the occupational choice.

5.2.1 Occupational Choice

Young adults choose their occupation to maximize utility over their working period of T years, subject to discount factor r, by solving:

$$\max_o \left\{ \int_0^T e^{-rt} \left(\log(C_{io}) + \tau_{ok} + A_o\right) dt \right\}, \tag{2}$$

where C_{io} is consumption, τ_{ok} are workers' preferences for working in their traditional occupation, and A_o are occupational amenities.Workers spend their entire income on the final consumption good (which is the numeraire), so that the budget constraint is equal to:

$$C_{io} = (1 - T_k) w_o \Theta_{io}, \tag{3}$$

where T_k is caste wage discrimination, w_o is the occupation-specific wage rate, and Θ_{io} are the total human capital units that a worker supplies to occupation o. This human capital measure depends on workers' own characteristics and their social environment and is given by:

$$\Theta_{io} = (\alpha_i \beta_i)^{\rho_o} \pi_{io} \psi_{io}, \tag{4}$$

where ψ_{io} captures productivity effects from workers' caste networks and parental occupation, π_{io} are occupation-specific skills, α_i, β_i are measures of general ability,[17] and ρ_o captures occupation-specific returns to general ability. Substituting the budget constraint (Equation 3) and the expression for human capital (Equation 4) into the utility maximization (Equation 2) allows us to formulate the occupational choice problem as:

$$\max_o \left\{ \int_0^T e^{-rt} \left[\log \left((1 - T_k) \, w_o \, (\alpha_i \beta_i)^{\rho_o} \, \psi_{io} \right) + \tau_{ok} + A_o + \log(\pi_{io}) \right] dt \right\} \equiv \bar{r} \max_o \left\{ \bar{u}_{io} + \log(\pi_{io}) \right\},$$

where $\bar{r}$ captures the discount factor and $\bar{u}_{io}$ is the expected lifetime utility (net of the occupation-productivity shock) of choosing occupation o. We provide the definitions and derivations in Appendix A3.1.

Solving the Occupational Choice Problem:

To solve this discrete choice problem, we impose the following assumptions:

Assumption 1. Idiosyncratic productivity shocks $\log(\pi_{io})$ are i.i.d. across occupation choices and are distributed Type-I Extreme Value with zero mean: $\Pr(\epsilon \leq x) = \exp(-\exp(-x - \bar{\gamma}))$.

Assumption 2. Idiosyncratic ability α_i is i.i.d. across workers and is distributed log-normal with mean 0 and variance σ_α^2.

Under Assumption 1, we can express the probability that worker i with ability α_i chooses occupation o as:

$$P_{io|\alpha_i} = \frac{(\exp \bar{u}_{io})^{\sigma_\pi}}{\sum_{o'} (\exp \bar{u}_{io'})^{\sigma_\pi}}, \tag{5}$$

and the expected utility before knowing occupation-specific productivity shocks π_{io} as:

$$\mathbb{E}_{\pi_{io}} \left[\bar{r} \max_o \left\{ \bar{u}_{io} + \log(\pi_{io}) \right\} \right] = \frac{\bar{r}}{\sigma_\pi} \log \sum_o (\exp \bar{u}_{io})^{\sigma_\pi}, \tag{6}$$

where σ_π captures the dispersion of the idiosyncratic productivity shocks π_{io}. We then use Assumption 2 to integrate over unobservable ability α_i so that worker i's unconditional occupational choice

[17] Recall that α_i are unobserved ability shocks and β_i is the observed component of ability, which is determined by workers' education and experience. Since individuals choose their education during childhood, we can treat it as a fixed characteristic in the occupational choice problem.

probability is equal to:

$$P_{io} = \int P_{io|\alpha_i} \phi\left(\alpha_i\right) d\alpha_i,$$

where $\phi\left(\cdot\right)$ indicates the log-normal PDF. This final integral has no closed-form solution.

5.2.2 Education Choice

Caste affects education choices via two main channels: first, caste identity can make it more or less costly for individuals to accumulate human capital. Second, caste identity affects individuals' expected returns to education. If an individual believes that she is likely to enter her traditional occupation–in which the returns to education are low–then she will invest less ex-ante in acquiring education. We therefore model children's education choice using the standard Mincerian formulation augmented by caste-specific costs:

$$\max_{s_i} \left\{ \left(\frac{\bar{r}}{\sigma_\pi} \log \sum_o \left(\exp \bar{u}_{io}\right)^{\sigma_\pi} \right) - \left(\kappa_{1k} + \frac{\kappa_{2k}}{2} s_i + \eta_i \right) s_i \right\}. \tag{7}$$

The first term of this equation represents the net present value of expected lifetime utility (derived in Equation 6) before knowing occupation-specific productivity shocks π_{io} (which are realized only after education is completed). Utility $\bar{u}_{io}$ increases in years of schooling and captures expected returns to schooling due to higher wages during workers' years in the labor force. The second term of equation 7 represents the costs of education, including both caste-specific shifters κ_k and idiosyncratic education cost shocks η_i. To facilitate estimation of the κ_k parameters, we make the following assumption:

Assumption 3. Education cost shocks η_i are i.i.d. across workers and distributed Normal with mean 0 and variance σ_η^2.

When choosing the optimal amount of schooling, individuals weigh marginal costs against expected marginal returns of schooling. We can define the optimal schooling level implicitly by differentiating equation 7 with respect to s_i. We present the full derivations, including integration and optimization, in Appendix A3.2.

5.2.3 Aggregation of Human Capital in each Occupation

Individuals' education and occupational choices jointly determine schooling levels, the allocation of general and occupation-specific talent, and the structure of occupational caste networks. These factors together determine total human capital supply in each occupation, which we now derive. To do so, we first solve for worker's expected occupation-specific productivity π_{io} conditional on having chosen occupation o which is equal to:

$$\mathbb{E}\left(\pi_{io|\alpha_i}\right) = \sigma_\pi \left(\frac{1}{P_{io|\alpha_i}}\right)^{\frac{1}{\sigma_\pi}} \mathit{\Gamma}\left(1 - \frac{1}{\sigma_\pi}\right), \tag{8}$$

where $\Gamma(\cdot)$ is the gamma function and where we used Assumption 1 to derive the closed form solution. This expression illustrates the negative selection within α-type in the uncorrelated Roy model: if traditional occupation affinity increases $P_{io|\alpha_i}$, then expected human capital must decrease. Next, we combine this measure with workers other characteristics (cf. Equation 4) to express the total expected human capital units in occupation o as:

$$\mathbb{E}\left(\Theta_o\right) = \sum_i \int_{\alpha_i} P_{io|\alpha_i} \left(\alpha_i \beta_i\right)^{\rho_o} \psi_{io} \mathbb{E}\left(\pi_{io|\alpha_i}\right) d\phi\left(\alpha_i\right), \tag{9}$$

where we weigh each observation by the corresponding occupational choice probability and use Assumption 2 to integrate unobservable ability α_i over the PDF of the Normal distribution, $\phi(\cdot)$. Across α-types, the negative relationship between traditional occupation and human capital is no longer guaranteed. If the utility of working in the caste occupation sufficiently increases $P_{io|\alpha}$ for high α-types in low returns traditional occupations, then the average human capital of traditional workers will be greater than that of outsiders due better α-skill composition.

5.2.4 Social Networks

We define caste-occupation networks as the share of a worker's occupation that consists of members from their caste, so that networks are endogenously determined by occupational choices. When constructing counterfactual social networks, given occupational choice probabilities $P_{io|\alpha_i}$, we calculate networks as,

$$\text{SocialNetwork}_{ok} = \frac{\sum_{i \in k} \int_{\alpha} P_{io|\alpha_i} d\phi\left(\alpha_i\right)}{\sum_i \int_{\alpha} P_{io|\alpha_i} d\phi\left(\alpha_i\right)}$$

The productivity effects of social networks imply that workers' occupational choices have important externalities on their fellow-caste members. In the absence of a coordinating mechanism, individuals do not internalize these effects and equilibrium social networks may feature less clustering of castes into the same occupation than in an output-maximizing allocation.

5.3 Firms and Market Clearing

Perfectly competitive firms produce the final consumption good C. The production technology is CES and uses human capital units from each occupation Θ_o as inputs. Profit maximization is therefore given by:

$$\max_{\Theta_o} \left\{ A \left[\sum_{o'} Z_{o'} \Theta_{o'}^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}} - \sum_o w_o \Theta_o \right\}, \tag{10}$$

where A is total factor productivity, Z_o is the factor share of each occupation's human capital and σ is the elasticity of substitution between occupations. Firms' first order condition (with respect to Θ_o) determine the demand for human capital in each occupation. Wage rates w_o adjust in equilibrium to ensure that labor markets clear, equalizing human capital demand and human capital supply (cf. Equation 9) in each occupation.

5.4 Equilibrium

We formally define the equilibrium in Appendix A3.3. The Appendix further describes how we endogenize wage discrimination T_k by assuming that entrepreneurs derive a disutility from employment workers from certain castes (similar to Hsieh et al. (2013)).

5.5

6 Structural Estimation

We use maximum likelihood to directly match the model's expressions for individuals' wages, education, and occupation choices to their counterparts in our individual-level data. Consistent with the timing assumptions in our model, we estimate two separate likelihood functions: the first for the probability of the observed occupational choices and wages, and the second for the likelihood of the observed educational choices. We first describe the parametrization of model components and then present the likelihood functions.

6.1 Parameterization and Heterogeneous Effects

Preferences for traditional occupation: The non-pecuniary utility that individuals' receive from an occupation is allowed to vary according to their traditional occupation and the hierarchical rank of their caste:

$$\begin{aligned}\tau_{ok} &= \mathbb{I}\left(\text{TraditionalOccupation}_k = o\right)\left(\tilde{\tau}_1 + \tilde{\tau}_2\mathbb{I}\left(\text{OBC}_k\right) + \tilde{\tau}_3\mathbb{I}\left(\text{SC}_k\right) + \tilde{\tau}_4\mathbb{I}\left(\text{ST}_k\right) + \tilde{\tau}_5\mathbb{I}\left(\text{Female}_i\right)\right)\\ &+ \mathbb{I}\left(\text{Homework}_o\right) \times \mathbb{I}\left(\text{Female}_i\right) \times \left(\tilde{\tau}_6 + \tilde{\tau}_7\mathbb{I}\left(\text{OBC}_k\right) + \tilde{\tau}_8\mathbb{I}\left(\text{SC}_k\right) + \tilde{\tau}_9\mathbb{I}\left(\text{ST}_k\right)\right),\end{aligned}$$

where $\mathbb{I}\left(\text{TraditionalOccupation}_k = o\right)$ indicates whether occupation o is a traditional occupation of caste k. We further allow the value of τ_{ok} to differ depending on castes' social ranking–captured by the indicator variables $\mathbb{I}\left(\text{OBC}_k\right), \mathbb{I}\left(\text{SC}_k\right), \mathbb{I}\left(\text{ST}_k\right)$–to accommodate the fact that many traditional occupations of lower castes might not convey positive utility (for example, carcass removal).[18] In addition, women may face social sanctions when working outside the home, so that we include an indicator for homework, interacted with a gender dummy. We again allow this effect to vary according to the caste hierarchy, following evidence in Eswaran et al. (2013) and Cassan and Vandewalle (2020) showing that this stigma is strongest for high castes women.

[18] Father's occupation might also have a direct impact on utility and could therefore be included in τ_{ok}. We tested this alternative specifications and found this effect to be insignificant for males and slightly negative for women, perhaps due to differing gender roles. In our preferred specification, we therefore allow for effects from father's occupation only through the productivity shifter ψ_{io}. This choice also offers a cleaner identification, since simultaneous effects of parental occupation on their children's utility *and* productivity could only be separately identified by functional form.

Observable components of general ability: Using the standard Mincer formulation, we parameterize observable human capital, β_i as a function of education s_i, experience and experience squared:

$$\begin{aligned}\beta_i = \exp\big(&\tilde{\beta}_1 \text{experience}_i + \tilde{\beta}_2 \text{experience}_i^2 + \tilde{\beta}_3 s_i + \\ &+ \mathbb{I}\left(\text{Female}_i\right)\big(\tilde{\beta}_4 \text{experience}_i + \tilde{\beta}_5 \text{experience}_i^2 + \tilde{\beta}_6 s_i\big)\big).\end{aligned}$$

where we define experience as $age_i - s_i - 6$.

Productivity effects from social environment: We model productivity effects from caste networks and father's occupation as:

$$\begin{aligned}\psi_{io} = \exp\big(&\tilde{\psi}_1 \mathbb{I}\left(\text{Father occ} = o\right) + \tilde{\psi}_2 \text{SocialNetwork}_{ok} + \\ &+ \mathbb{I}\left(\text{Female}_i\right)\big(\tilde{\psi}_3 \mathbb{I}\left(\text{Father occ} = o\right) + \tilde{\psi}_4 \text{SocialNetwork}_{ok}\big)\big),\end{aligned}$$

where $\text{SocialNetwork}_{ok}$ is the share of all workers in occupation o that are members of caste k[19] and $\mathbb{I}\left(\text{Father occ} = 0\right)$ indicates whether the father of individual i worked in occupation o. We allow the effects of caste networks and parental occupation to differ for women, reflecting the fact that fathers may differentially pass down their occupation-specific knowledge to sons or daughters, and that social networks may be less important for women (as shown in (Munshi and Rosenzweig, 2006)). We set $\text{SocialNetwork}_{ok} = 0$ for homework, since social networks do not seem relevant in this setting.

Wage discrimination:

We parametrize wage discrimination based on caste hierarchy and gender in the following way:

$$(1 - T_k) = \exp\big(\tilde{\delta}_1 \mathbb{I}\left(\text{Female}_i\right) + \tilde{\delta}_2 \mathbb{I}\left(\text{OBC}_k\right) + \tilde{\delta}_3 \mathbb{I}\left(\text{SC}_k\right) + \tilde{\delta}_4 \mathbb{I}\left(\text{ST}_k\right)\big).$$

We set $T_k = 0$ for homework, reflecting the fact that caste discrimination within the home is unlikely, and that women may face systematic discrimination in all market occupations.

Cost of schooling: We allow education cost per year of schooling κ_k to vary by caste hierarchy, gender, and the number of years of schooling,

$$\begin{aligned}\kappa_k \quad &= \quad \tilde{\kappa}_1 \mathbb{I}\left(\text{Female}_i\right) + \tilde{\kappa}_2 \mathbb{I}\left(\text{OBC}_k\right) + \tilde{\kappa}_3 \mathbb{I}\left(\text{SC}_k\right) + \tilde{\kappa}_4 \mathbb{I}\left(\text{ST}_k\right) \\ &+ \quad \text{YearsEducation}_i \times \left(\tilde{\kappa}_5 \mathbb{I}\left(\text{Female}_i\right) + \tilde{\kappa}_6 \mathbb{I}\left(\text{OBC}_k\right) + \tilde{\kappa}_7 \mathbb{I}\left(\text{SC}_k\right) + \tilde{\kappa}_8 \mathbb{I}\left(\text{ST}_k\right)\right).\end{aligned}$$

Occupation Parameters: At the occupation level, we estimate amenities A_o, wage rates w_o, and skill returns ρ_o. These variables are simple vectors whose elements represent occupational categories.

Distribution Parameters: Last, we need to estimate the dispersion of the three idiosyncratic shocks in our model. First, for occupation-specific productivity shocks π_{io} which are extreme

[19] We jacknife this variable for the individual's own occupation, subtracting 1 from both the number of caste members and the total workers in the occupation.

value distributed with dispersion parameter σ_π; second, for general ability shocks α, which are log-normally distributed with mean 1 and standard deviation σ_α; and third for education cost shocks η_i, which are normally distributed with mean 0 and standard deviation σ_η.

6.2 Likelihood Function

With these parametrizations at hand, we now turn to our maximum likelihood estimation. The estimation proceeds in two steps: the first for occupation and wages, and the second for education.

Occupation and wage likelihood

The occupation and wage likelihood function estimates the first set of parameters, $\Omega_{occ} = \Big\{\tilde{\tau}, \tilde{\beta}, \tilde{\psi}, \tilde{\delta}, A_o, w_o, \rho_o, \sigma_\alpha, \sigma$
The likelihood that a worker i and earns wage y_{io} in occupation o can be expressed as the product of the probability that she chooses occupation o and the probability that she earns wage y_{io} conditional on that occupational choice:

$$L_i\left(\hat{y}_{i\hat{o}}, \hat{o}; \Omega, X_i\right) = \int_\alpha \Pr\left[y_{io} = \hat{y}_{i\hat{o}} | o = \hat{o}; \Omega, X_i, \alpha\right] \times \Pr\left[o = \hat{o} | \Omega, X_i, \alpha\right] d\alpha, \tag{11}$$

where X_i, $\hat{o}$ and $\hat{y}_{i\hat{o}}$ are individual characteristics, chosen occupation, and realized wages in this occupation, as observed in the data. Under the assumption of extreme value distributed productivity shocks, the model admits a closed form expression for occupational choice probability (as derived above in Equation 5):

$$\Pr\left[o = \hat{o} | \Omega, X_i, \alpha_i\right] = P_{io|\alpha_i}$$

and for the conditional wage probability (derived in Appendix A3.4):

$$\begin{aligned}\Pr\left[y_{io} = \hat{y}_{i\hat{o}} | o = \hat{o}; \Omega, X_i, \alpha_i\right] \quad = \quad & \frac{\sigma_\pi}{\hat{y}_{i\hat{o}}} \left(\frac{\sum_{o'} \left(\exp \bar{u}_{io'|\alpha_i}\right)^{\sigma_\pi}}{\left(\exp\left(\tau_{ok} + A_o + \rho_o\left(\bar{\beta} - \beta_i\right)\right) \hat{y}_{i\hat{o}}\right)^{\sigma_\pi}} \right) \times \\ & \exp\left\{ - \left(\frac{\sum_{o'} \left(\exp \bar{u}_{io'|\alpha_i}\right)^{\sigma_\pi}}{\left(\exp\left(\tau_{ok} + A_o + \rho_o\left(\bar{\beta} - \beta_i\right)\right) \hat{y}_{i\hat{o}}\right)^{\sigma_\pi}} \right) \right\}.\end{aligned}$$

This component of the likelihood is not defined for home workers since they have no observable wage data. We therefore set it to 1 for these individuals, using only the occupational choice data to estimate homework occupational parameters. In our counterfactuals, we do not endogenize the homework "wage", and present results on output only for market workers.

Both of these probabilities are conditional on the level of unobserved ability α_i, over which we integrate in the likelihood function in Equation 11.[20] In the estimation, we normalize the mean and standard deviation of the general α skill distribution for males to one. This normalization does not affect the likelihood value since the mean of unobserved skills is not separately identified from the average wage rates w_o, and the variance of unobserved skills is not separately identified from the scale of skill returns ρ_o and the coefficients $\tilde{\beta}$. More generally, an environment with high returns to

[20] Specifically, we integrate over α using Gauss-Hermite quadrature with 7 nodes.

skill in all occupations and a small variance of skills is observationally equivalent to cases with low returns to skill and a large variance of skills. We normalize the occupational amenity A_o in the first occupational category to 1, since occupational utilities are only identified in relative terms. Finally, we normalize w_o for homework to 1: without wage data for home workers this is not separately identified from A_o.

Education likelihood

With the first set of parameters at hand, we now use the education likelihood to estimate the remaining parameters: $\Omega_{edu} = \{\tilde{\kappa}, \sigma_\eta\}$. We specify the education likelihood function as a Tobit to account for the fact that almost a third of individuals in the data have no formal education. For these individuals (with $\hat{s}_i = 0$) the schooling choice is likely inframarginal, so the education likelihood corresponds to the probability that their education cost shocks are sufficiently high to censor their education levels at zero. The full education likelihood is therefore equal to:

$$L_i\left(\hat{s}_i\right) = \int_\alpha \left(\frac{1}{\sigma_\eta}\phi\left(\frac{\hat{\eta}_{i\alpha}}{\sigma_\eta}\right)\right)^{\mathbb{I}(\hat{s}_i>0)} \left(1-\Phi\left(\frac{\hat{\eta}_{i\alpha}}{\sigma}\right)\right)^{\mathbb{I}(\hat{s}_i=0)} d\alpha, \tag{12}$$

where ϕ is the PDF and Φ the CDF of the standard normal distribution. $\hat{\eta}_{i\alpha}\left(\hat{s}_i, \hat{y}_{io}, \hat{o}; \Omega\right)$ are individuals' education cost shocks which rationalize observed education choices $\hat{s}_i$, conditional on individuals' observed wages $\hat{y}_{io}$, observed occupations $\hat{o}$ and parameters Ω. We can characterize individuals' education cost shocks from the first order conditions of the education choice problem (Equation 7) in the following way:

$$\hat{\eta}_{i\alpha} = -\kappa_{1g} - \kappa_{2g}s_i + \bar{r}\left[-\frac{r}{\sigma_\pi}\log\sum_o \exp\left(\sigma_\pi \bar{u}_{io}\right) + \tilde{\beta}_s \sum_o \rho_o P_{io|\alpha_i}\right],$$

where the term in brackets represents expected returns to education during individuals' working period. We provide the full derivation of this expression in Appendix A3.2. As before, we integrate each likelihood contribution in Equation 12 over the distribution of possible ability shocks α.

We solve for the likelihood subject to the constraint that the second-order conditions of the education choice problem (Equation 7) are negative at the optimal education level to ensure that our retrieved education choices maximize agents' utility.[21]

It is theoretically possible to estimate the occupation/wage and education likelihoods simultaneously. However, it would be computationally infeasible to impose the second order constraint from the education choice problem on the combined likelihood , since the constraint is linear in κ for the education likelihood, but non-linear in most other parameters. We therefore implement the estimation in two steps and we bootstrap the standard errors to account for this 2-stage process, clustering at the PSU level.

[21] The education choices is not well defined in 55 out of 688,380 individual×alpha-type combinations as our simulated education choices correspond to local rather than global maxima of the utility function. We interpret this as rejections of the possibility of observing these α values for these particular individuals, so that we drop them from the estimation.

6.3 Backing out production parameters

Last, we need to determine the parameters from the CES production technology, i.e. occupational intensity Z_o, total factor productivity A, and the elasticity of substitution σ across occupations. Following the literature, we set σ equal to 2/3. We compute these parameters by matching the model's optimality conditions to the data conditional on our previous parameter estimates Ω. In particular, dividing firms' first order conditions across two occupations yields the following expression:

$$\frac{Z_o}{Z_{o'}} = \frac{w_{o'}}{w_o} \left(\frac{\Theta_o}{\Theta_{o'}} \right)^{\frac{-1}{\sigma}}, \tag{13}$$

which shows that we can compute relative occupational intensities Z_o by using our estimated wage rates and by constructing human capital supply in each occupation from the data. Factor shares Z_o have to sum to one across all occupations so that all values of Z_o are pinned down by Equation 13.Conditional on knowing Z_o, w_o, Θ_o and σ, we can again use firm's first order conditions to infer total factor productivity A in the following way:

$$A = \frac{w_o}{Z_o \Theta_o^{\frac{-1}{\sigma}} \left[\sum_o Z_o \Theta_o^{\frac{\sigma-1}{\sigma}} \right]^{\frac{1}{\sigma-1}}}. \tag{14}$$

7 Results

7.1 Structural Parameters

We present our maximum likelihood estimates in Tables 5 and 6.

Our estimates of the non-pecuniary utility of working in one's traditional occupation–the key focus of our paper–are displayed in the first two columns of Table 5. To provide an interpretation of the estimated parameter values, let us consider their effects on occupational choice probabilities, which are shifted by $\exp(\sigma_\pi \tau_{io})$ (cf. Equation 5). The coefficient on traditional occupation, 0.178, implies that general caste men[22] choose their traditional occupation with a 23.1 percent greater probability than non-traditional occupations. This effect is smaller for women–combining the coefficients in rows 1 and 2 implies only a 12.4 percent greater attraction to the traditional occupation for female workers. Another perspective comes from comparing the τ_{io} parameters to the variation in occupational amenities A_o, the other non-pecuniary source of occupational utility. Here we see that the τ_{io} shifters are relatively small: the standard deviation of amenities A_o is 2.98 times larger than men's preferences for their traditional occupation. Relative to general castes, the appeal of the traditional occupation is significantly stronger for OBC castes, not significantly different for scheduled castes, and significantly weaker for the scheduled tribes. In rows 6-9 of Table 5 we see that women have a very strong affinity for homework (or, equivalently, stigma for market work). This effect is weaker for castes of lower social status as suggested by the literature.

The second pair of columns of Table 5 displays the estimated coefficients $\tilde{\beta}$ that transform years

[22] A caste is "general" if it is neither SC, ST nor OBC.

of education and experience into general human capital units β_i. The values displayed here are the baseline returns; for any specific occupation, the actual returns are adjusted by the occupation-specific return to general human capital ρ_o (shown in Table A2). The coefficient on years of schooling is analogous to the standard Mincer coefficient and, averaging the adjusted returns over occupations, our estimates of 0.10 for men and 0.09 for women are in line with other studies that focus on Indian context or similar countries (Psacharopoulos and Patrinos, 2004). Returns to experience are increasing, with diminishing returns, for men, while experience has little or negative returns for women.

Column group 4 in Table 5 shows the estimated parameters $\tilde{\psi}$ that determine the productivity effects from caste-occupation networks and from working in the same occupation as one's father ψ_{io}. We find very strong intergenerational effects: individuals working in their father's occupation earn on average 58 percent more compared to other workers in the same occupation. We also find very strong network effects: a 1% increase in the share of workers of an occupation who are of the same caste as the respondent is associated with a 8% increase in wages. Consistent with the reduced form effects in Table 3, these effects seem to be even stronger for women.

The fifth set of columns displays the parameters that determine castes' wage discrimination, $(1 - T_k)$. We find that the main victims of discrimination are women, who earn only 33.2 percent of the wages from identical men in the same occupation. We do not find wage discrimination against OBC and SC castes–indeed wages appear marginally higher for both groups and scheduled tribes show only a small negative discrimination effect. Since these castes benefit from positive affirmative action in many professions (e.g. through quotas in university and medical school admissions), it is likely that our estimates reflect the combined effect of negative discrimination and positive social policies.

Structural coefficients that determine the cost of education are presented in column group 3. We find that these costs are convex and negative for low years of education with costs first decreasing and then increasing after around 6 years of schooling. Costs become positive for schooling levels beyond 9 years for women and beyond 12 years for men.[23] Costs ultimately rise more steeply for SC and ST castes. These negative non-pecuniary costs of receiving low education can reflect social stigma, returns on the marriage markets, or compulsory elementary schooling. Recall that schooling choices also depend on idiosyncratic education cost η_i as well as forgone earnings.

Occupation characteristics: For each occupation, we separately estimate wage rates w_o, amenities A_o, and returns to general human capital ρ_o. For conciseness we display the full vectors of parameters in Appendix Table A2. For each occupation, the wage rate per human capital unit can be interpreted as the intercept of the wage function for individuals with very low human capital. Consistent with this interpretation, the occupations with the highest wage rates ($\ln w_o$) are construction (0.95), and agricultural labor (0.87), and the ones with the lowest values are legal professionals (-9.21) and doctors (-8.41). The highest occupational amenities ($\ln A_o$) are for animal

[23] See Figure A1 for a graphical presentation of the κ values.

farmers (3.95) and non-labor income earners (3.71), while garbage workers (1.79) and plantation workers (1.93) have the lowest. Finally, the returns to general human capital (ρ_o) are highest for professors/teachers (1.29) and legal professionals (1.14), and lowest for makers of tobacco products (-0.53) and animal farmers (-0.52). We consider it re-assuring that many of these estimates are consistent with reasonable beliefs about the nature of different occupations.

7.2 Counterfactual Results

Our counterfactual analysis explores how the Indian economy would differ if caste identity was not linked to traditional occupations. Importantly, our model features three channels that can lead to an over representation of castes in their traditional occupations. First, the direct attachment to traditional occupation τ_{io}–which we eliminate in all of our counterfactuals, formally setting $\tilde{\tau}_1 = \tilde{\tau}_2 = \tilde{\tau}_3 = \tilde{\tau}_4 = \tilde{\tau}_5 = 0$. Second, workers are more productive when they work in the same occupation of their father. To analyze the importance of this channel, we consider a counterfactual in which there is no correlation between traditional occupationand the distribution of father's occupation . We do so by regressing an indicator for father's occupation on an indicator for traditional occupation and a constant in a dataset at the occupation×individual level. We then replace fathers' observed occupations with the residual from this regression–which is by definition orthogonal to traditional occupations. To avoid mechanical effects on aggregate output, we rescale the residual to the same mean as the original father-occupation-indicator. Third, workers productivity increases in the size of their caste network in their chosen occupation. Once networks are established, this effect can sustain the selection of caste members into their traditional occupations. To evaluate the importance of this mechanism, we first implement counterfactuals in which we hold caste-occupation networks fixed and then we allow networks to adjust endogenously. We assume that individuals have perfect information about all observable variables including wages and occupation-caste networks.[24]

We present our results in Table 7. All three counterfactuals in Panel (A) hold caste-occupation networks constant. Columns 1 and 2 also hold fathers' occupational distribution constant, so that these counterfactuals answer the question, "what would happen if the current generation of workers stopped valuing their traditional occupation?" Column 1 shows the direct effect of eliminating traditional occupation preference τ_{io} when we only allow occupational choices and wages to adjust endogenously, holding everything else constant. The impact on the aggregate economy is extremely minor: output increases by 0.06 percent and human capital by 0.05 percent. In addition, labor force participation decreases–primarily due to women leaving traditional occupations and entering homework–so that output per worker increases slightly more (0.33 percent). In column 2, we also

[24] Our counterfactuals use the posterior distribution of α_i and the values of η_i generated during the estimation. Thus when simulating occupation choices, years of schooling, and wages at the estimated parameters we generate a baseline very close to the empirical values of these outcomes. We maintain these vectors of (α_i, η_i) unobservables when considering alternative parameters, assuming that the value of η_i for individuals with zero education is the mean of the set of η_i's consistent with this choice. Alternatively, using the generic normal distributions for α_i and η_i yields nearly identical counterfactual implications, but with greater computational burden.

allow individuals to adjust their education choices, which increases schooling by 0.49 percent and output by 0.35 percent–again overall effects are small.

Why is the direct impact of removing traditional occupation affinity so small? First, as mentioned above, the magnitude of the $\tilde{\tau}$ parameters is small relative to the variation in other structural parameters, in particular relative to amenities A_o and productivity effects from networks and fathers' occupation ψ_{io}. Thus the basic structure of the economy is relatively unchanged as shown in Panel A.ii of Table 7: total employment shares drop by at most 0.72 percentage points for any given occupation, even if the share of traditional workers drop by up to 5.35 percentage points for the most affected occupation. This finding implies that traditional workers simply get replaced by other similar workers, thus keeping the occupational structure and aggregate output similar. Despite the trivial aggregate effects, we do find that removing preferences for traditional occupations reduces the share of workers who work in their traditional occupation from 10.4 to 9.2 percent–a reduction of 11.6 percent. In addition, we see improvements in workers' selection based on their individual characteristics–their occupation-specific productivity π_{io} and their general ability α_i and β_i, as individuals with higher general ability increasingly select into occupations with high returns to ability ρ_o.[25] However, these gains are partially offset by the fact that workers select less into their traditional occupation–and consequentially also less into their father's occupation–which reduces productivity gains from caste networks and intergenerational transfers ψ_{io}.

In column 3 of Table 7 we remove one more channel of historical persistence by removing the correlation between father's occupationand traditional occupations, as described above. Compared to the baseline, aggregate output now decreases by 2.98 percent. A part of this drop is driven by a (3 percent) decline in labor force participation as some workers who were previously attracted to occupations that were simultaneously their father's and their traditional occupation now choose home production. Output per worker decreases slightly (by 0.005 percent) due to the reduced complementarity between parental occupation and occupational networks. In particular, by eliminating the correlation between father's and traditional occupations, we now make individuals choose between either their father's occupation or their traditional occupation where caste networks (which we hold constant) are strongest. In the aggregate, productivity reductions due to smaller caste networks and less intergenerational transfers more than offset improvements in selection based on individual characteristics. Removing these complementarities has large distributional effects: the most affected caste sees a 53.5 percent decrease in income, the most affected occupation looses 20.6 percent of its human capital, and the aggregate share of workers who work in their traditional occupations decreases by 31.6 percent (from 10.4 to 7.1 percent).

In Panel B of Table 7, we implement the same counterfactuals, now allowing caste networks to adjust endogenously. For this analysis, we first solve for the baseline steady state by fixing all parameters at their estimated values and by iterating over caste-occupation networks and occupational human capital until these objects are consistent with individuals' choices. We then compare all counterfactuals in Panel B to this baseline steady state. There is a possibility of

[25] Recall that we model general ability through unobserved general ability shocks α_o and observed ability β_i which depends on schooling and experience.

multiple equilibria due to the productivity spillovers from caste networks. While we have not exhaustively investigated the set of all possible equilibria, we adopt a procedure designed to identify the equilibrium that is plausibly "closest" to the existing one.[26] We argue that this equilibrium best captures how caste-occupation networks might evolve if attachment to traditional occupations would disappear. As a robustness check, we implement all counterfactuals with weaker network effects to assess the sensitivity of our results to these parameter estimates (see Appendix A4.2 for more detail).

Column 1 of Panel B in Table 7 shows that the removal of traditional occupation affinity with endogenous networks leads to a very small reduction in output (0.145 percent). This drop is driven by a decrease in labor market participation (-0.3 percent), so that output per worker increases slightly by 0.17 percent. In column 2, we allow schooling to adjust endogenously, which leads to a 0.76 percent increase in output and a 1.1 percent increase in output per worker. Compared to Panel A, gains are now larger with endogenous networks, because workers reallocate their human capital more efficiently and choose more education. The share of traditional workers decreases by 3.86 percentage points in the most affected occupation and by 6.17 percentage points in the most affected caste. The aggregate share of workers who work in their traditional occupation decreases by roughly 9.5 percent.

Finally, in column 3 we eliminate the influence of traditional occupation via parental occupation. Effects are now remarkably large with a 8 percent decrease in output, a 3.2 percent decline in labor force participation and a 5 percent drop in output per worker. These results highlight the important coordinating role played by castes' affinity for their traditional occupation: due to social network effects, individuals' occupational choices have a large externality and traditional occupations, either directly via preferences or indirectly via parental occupation, organize castes into strong occupational social networks. Without this coordinating element, the resulting occupational networks are therefore only weakly clustered at the caste level. Consistent with this, we find that the share of traditional workers decreases by 4 percentage points (24 percent) in the aggregate, by 10.6 percentage points in the most affected occupation and by 27.5 percentage points in the most affected caste. The drop in output due to weaker social networks and less parental learning is somewhat offset by a larger increase in education (by 2 percent) and improved selection based on individual characteristics–without strong caste networks in low-returns occupations, agents invest more in education and are more likely to pursue occupations aligned with their comparative advantage.

To document more systemically how each counterfactual affects inequality, we compare the dispersion of human capital in the baseline and counterfactual. We focus on human capital as a measure of inequality because, unlike income, it is defined regardless of labor force participation. To do so, we use Growth Incidence Curves, proposed by Ravaillon and Chen (2003) and popularized by Milanovic (2016). We rank the population by percentile of their baseline human capital, and compute the mean growth rate of these percentiles.[27]

[26] To do this, we start from the estimated human capital distribution and caste networks and we perform only minor updates–changing exogenous parameters in several small steps–to converge to the new equilibrium.

[27] Note that computing the mean growth rate is different from computing the growth rate of the mean, and has

The results, presented in Figure 2, show that most counterfactuals removing caste-occupation links have the largest positive effects for those in the middle of the baseline human capital distribution, and the most negative effects for individuals with low human capital. Figure 2a, showing effects when social networks remain as estimated, displays a particularly clear U-shaped pattern. The lowest human capital workers, particularly women who face labor force discrimination, have such low probability of entering high skill occupations that the counterfactuals do not induce them to invest in additional education. Meanwhile, they suffer from the reduction of paternal human capital in traditional (low-skill) occupation. The highest human capital workers are already in high skills sectors, and have little scope to increase education (since its costs are convex). Figure 2b, showing counterfactuals with endogenous social networks, displays a similar pattern. Decreases in human capital (including social network-based productivity spillovers) are particularly sharp for low-skill workers, who tend to work in traditional occupations with strong baseline social networks. High skill workers now benefit more, as social networks become stronger in their high skill occupations. These results confirm the findings of Munshi and Rosenzweig (2006) who show that castes strongly rely on social networks–especially low skill individuals in traditional occupations.

8 Conclusion

The effect of social identity upon occupational choice has often been highlighted as a potential distortion of human capital allocation and source of economic inefficiency. Examining this question in the context of the Indian caste system, we find mixed evidence for these claims. Occupational identity does have a major effect upon career choice–in India certain occupations are composed primarily of individuals born into that occupation, and the average person is more than 3 times as likely to enter their traditional occupation than any other. Thus traditional caste-based occupational identities continue to affect the selection of occupations well into the modern era.

However, these large effects on occupational choices have only a small effect on the overall efficiency of the economy. Three forces act to diminish the distortions of caste-based choices on the macro-economy. First, we find that working in a parents' occupation increases productivity and castes' traditional occupations are highly correlated with their parental occupations. Hence, even if the non-pecuniary utility of the traditional occupation is removed, many individuals continue to work in it because they have learned occupation skills from their fathers. Second, the clustering of castes into occupations generates a positive social network effect that partially compensates for the misallocation of human capital. Though individuals may be working in the "wrong" occupations, by doing so they increase the productivity of all their caste-mates in that occupation. Finally, the misallocation of human capital caused by traditional occupations is primarily limited to the reallocation of low-skilled individuals between low-skilled occupations. Since these individuals are numerous, the magnitudes appear high, but the strength of caste identity is not sufficient to draw many extremely high skilled individuals out of the "modern workforce".

preferable properties as discussed in Ravaillon and Chen (2003).

Our results are larger, and more negative, when we consider how social networks might adjust to a reduction of the links between caste and occupation. Once the ties of occupational affinity and parentally transmitted human capital are broken, we find that individuals' occupational choices create social networks that are less clustered at the caste level, and with lower network-based productivity spillovers. This causes an overall reduction in market output, despite smaller increases in education and improved human capital allocation. These results highlight the importance of taking a broad view of economic importance of frictions. While the direct effects of occupational preferences are small, their indirect effects via parental human capital and social networks are much more important.

An important limitation of this study, inherent in the revealed preference approach to occupational identity, is that we cannot identify the absolute value of occupation-linked utility, only the relative value. Thus we cannot distinguish between a scenario in which individuals receive positive utility from doing their traditional occupation, versus an alternative in which they receive negative utility from any other occupation. With this caveat in mind, we limit our counterfactual analysis to the analysis of income and inequality, leaving the study of overall welfare impacts to future study using a different methodology.

Our analysis suggests a possible explanation for the remarkable persistence of the occupational identities in the 21st century. If the static economic costs are mild, but individuals receive substantial utility from conforming with social norms, then it is likely that these norms may persist over long periods. This may be one reason why caste occupational identities have endured despite deep changes in the economic structure of the country which many thought would lead to the weakening of the caste system (Srinivas, 2003). Our analysis suggests that, as Ambedkar (1936) anticipated, the main costs of identity frictions may be dynamic and occur over the course of structural transformation. For individuals attached to disappearing occupations, whether Indian handloom weavers at the turn of the 20th century or American manufacturing workers at the beginning of the 21st, there may be substantial costs to occupational change. We leave the study of these important dynamics to future research.

References

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017). When should you adjust standard errors for clustering? *Working Paper*.

Akerlof, G. (1976). The economics of caste and of the rat race and other woeful tales. *The Quarterly Journal of Economics 90*(4), 599–617.

Akerlof, G. A. (1980). A theory of social custom, of which unemployment may be one consequence. *The Quarterly Journal of Economics 94*(4), 749–775.

Akerlof, G. A. and R. E. Kranton (2000). Economics and identity. *The Quarterly Journal of Economics 115*(3), 715–753.

Altonji, J. G. and T. A. Dunn (2000). An intergenerational model of wages, hours, and earnings. *The Journal of Human Resources 35*(2), 221–258.

Alvarez-Cuadrado, F., F. Amodio, and M. Poschke (2019, May). Selection and absolute advantage in farming and entrepreneurship: Microeconomic evidence and macroeconomic implications. Technical report.

Ambedkar, B. R. (1936). *Annihilation of Caste.* Speech prepared for the annual conference of the Jat-Pat-Todak Mandal of Lahore but not delivered.

Atkin, D., E. Colson-Sihra, and M. Shayo (Forthcoming). How do we choose our identity? a revealed preference approach using food consumption. *Journal of Political Economics*.

Benjamin, D. J., J. J. Choi, and G. Fisher (2016). Religious identity and economic behavior. *Review of Economics and Statistics 98*(4), 617–637.

Borkotoky, K., S. Unisa, and A. K. Gupta (2015). Intergenerational transmission of education in India: evidence from a nationwide survey. *International Journal of Population Research 2015.*

Bryan, G. and M. Morten (2018). The aggregate productivity effects of internal migration: Evidence from indonesia. *Journal of Political Economy Forthcoming.*

Cassan, G. (2019). Affirmative action, education and gender: Evidence from india. *Journal of Development Economics 136*, 51–70.

Cassan, G. and L. Vandewalle (2020). Identities and public policies: Unexpected effects of political reservations for women in india. *Working Paper*.

Chen, R. and Y. Chen (2011, October). The potential of social identity for equilibrium selection. *American Economic Review 101*(6), 2562–89.

Cohn, A., M. A. Maréchal, and T. Noll (2015). Bad boys: How criminal identity salience affects rule violation. *The Review of Economic Studies 82*(4), 1289–1308.

Conlon, F. (1981). *The Census of India as a Source for Historical Study of Religion and Caste.* New Delhi:Manohar Publications.

Crooke, W. (1896). *The Tribes and Castes of the North Western Provinces and Oudh.* Calcutta: Office of the Superintendent of Government Printing.

Desai, S. and R. Vanneman (2015). India human development survey-ii (ihds-ii), 2011-12.

Desai, S., R. Vanneman, and National Council Of Applied Economic Research, New Delhi (2008). India human development survey (ihds), 2005.

Deshpande, R. and S. Palshikar (2008). Occupational mobility: How much does caste matter? *Economic and Political Weekly*, 61–70.

Dumont, L. (1970). *Homo Hierarchicus: The Caste System and Its Implications.* London: Weidenfeld & Nicolson.

Eckert, F. and M. Peters (2018). Spatial structural change. Technical report.

Eswaran, M., B. Ramaswami, and W. Wadhwa (2013). Status, caste, and the time allocation of women in rural india. *Economic Development and Cultural Change 61*(2), 311–333.

Gupta, D. (2000). *Interrogating caste: Understanding hierarchy and difference in Indian society.* Penguin Books India.

Headley, Z. (2013). Nommer la caste. ordre social et catégorie identitaire en inde contemporaine. *La Vie des idées.*

Heise, S. and T. Porzio (2019). Workers' home bias and spatial wage gaps. Technical report.

Hnatkovska, V., A. Lahiri, and S. B. Paul (2013). Breaking the caste barrier. *Journal of Human Resources 48*(2), 435–473.

Hsieh, C.-T., E. Hurst, C. Jones, and P. Klenow (2013, January). The allocation of talent and U.S. economic growth. Technical report.

Iversen, V., A. Krishna, and K. Sen (2017, 04). Rags to riches? intergenerational occupational mobility in India. *Economic and Political Weekly Vol. 52*(Issue No. 44).

Kitts, E. (1885). *A Compendium of the Castes and Tribes Found in India.* Bombay: Education Society Press, Byculla.

Kumar, S., A. Heath, and O. Heath (2002). Changing patterns of social mobility: Some trends over time. *Economic and Political Weekly 37*(40), 4091–4096.

Lagakos, D. and M. E. Waugh (2013). Selection, agriculture, and cross-country productivity differences. *American Economic Review 103*(2), 948–80.

Lanjouw, P. and N. Stern (1998). *Economic Development in Palanpur over Five Decades.* Oxford University Press.

Milanovic, B. (2016). *Global Inequality: A New Approach for the Age of Globalization.* Harvard University Press.

Munshi, K. (2019). Caste and the indian economy. *FJournal of Economic Literature.*

Munshi, K. and M. Rosenzweig (2006). Traditional institutions meet the modern world: Caste, gender, and schooling choice in a globalizing economy. *American Economic Review 96*(4), 1225–1252.

Munshi, K. and N. Wilson (2011). Identity, occupational choice, and mobility: Historical conditions and current decisions in the american midwest. *Working Paper*.

Oh, S. (2019). Does identity affect labor supply? Technical report.

Phelan, S. and E. A. Kinsella (2009). Occupational identity: Engaging socio-cultural perspectives. *Journal of Occupational Science 16*(2), 85–91.

Psacharopoulos, G. and H. A. Patrinos (2004). Returns to investment in education: a further update. *Education economics 12*(2), 111–134.

Ravaillon, M. and S. Chen (2003). Measuring pro-poor growth. *Economic Letters 78.*

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford economic papers 3*(2), 135–146.

Schwarz, H. (2010). *Constructing the Criminal Tribe in Colonial India: Acting like a Thief.* Wiley Blackwell.

Shayo, M. (2009). A model of social identity with an application to political economy: Nation, class, and redistribution. *American Political science review 103*(2), 147–174.

Singh, K. (1996). *Communities, Segments, Synonyms, Surnames and Titles*, Volume 8. Oxford University Press.

Skorikov, V. B. and F. W. Vondracek (2011). Occupational identity. In *Handbook of identity theory and research*, pp. 693–714. Springer.

Srinivas, M. N. (2003). An obituary on caste as a system. *Economic and Political Weekly 38*(5), 455–459.

Vaid, D. (2012). The caste-class association in india: An empirical analysis. *Asian Survey 52*(2), 395–422.

Vaid, D. (2014). Caste in contemporary india: Flexibility and persistence. *Annual Review of Sociology 40*, 391–410.

Wiser, W. H. (1936). *The Hindu Jajmani System.* Lucknow Publishing House.

Young, A. (2014). Structural transformation, the mismeasurement of productivity growth, and the cost disease of services. *American Economic Review 104*(11), 3635–67.

9 Tables.

Table 1: Traditional Occupation and Occupational Choice

	Probability of occupational choice			
	(1)	(2)	(3)	(4)
	A. Male (N =2,269,092)			
Occ. is caste's trad. occ.	0.068***	0.041***	0.038***	0.043***
	(0.002)	(0.002)	(0.002)	(0.002)
Occ. is father's occ.		0.307***	0.306***	0.306***
		(0.004)	(0.004)	(0.004)
Caste-occ. network			0.103***	0.105***
			(0.007)	(0.007)
Occ. is caste's trad. occ. * SC				-0.027***
				(0.004)
	B. Female (N =2,535,946)			
Occ. is caste's trad. occ.	0.027***	0.007***	0.006***	0.006***
	(0.002)	(0.002)	(0.002)	(0.002)
Occ. is father's occ.		0.140***	0.139***	0.139***
		(0.005)	(0.005)	(0.005)
Caste-occ. network			0.030***	0.030***
			(0.004)	(0.004)
Occ. is caste's trad. occ. * SC				-0.004
				(0.003)
Individual FE	Yes	Yes	Yes	Yes
Occ. FE	Yes	Yes	Yes	Yes

This table reports results of a linear probability model of occupational choice, using data from all 18-60 year old respondents of the 2011 IHDS. The dataset contains all unique combinations of respondent and occupation, with the outcome variable equal to 1 for the occupation chosen by the respondent, and 0 for all other occupations. The "Occ. is caste's trad. occ." variable indicates the occupations (if any) that are traditionally performed by the respondent's caste, as defined in section 3. The caste-occ. network data variable is equal to the jacknifed ratio of the number of members of the respondent's caste in an occupation to the total workers in the occupation. The "SC" variable indicates that the respondent's reported caste is part of the state-level list of scheduled castes. Standard errors clustered at the PSU (village) level. $*p < 0.10, **p < 0.05, ***p < 0.01$

Table 2: Traditional Occupation and Wages

	Log wages in chosen occupation			
	Male		Female	
	(1)	(2)	(3)	(4)
Occ. is own caste's	-0.238***	0.128***	-0.191**	0.115**
trad. occ.	(0.041)	(0.044)	(0.078)	(0.045)
Occ. is father's occ.	-0.046	0.062***	0.198	0.216***
	(0.076)	(0.019)	(0.171)	(0.068)
Caste-occ. network	1.726***	0.498*	3.654***	2.779**
	(0.462)	(0.249)	(1.052)	(1.183)
Occ. is caste's	0.172***	-0.065	0.317**	-0.083***
trad. occ.×SC	(0.058)	(0.049)	(0.157)	(0.026)
Jati FE	Yes	No	Yes	No
Occ. FE	No	Yes	No	Yes
R-sq	0.204	0.254	0.177	0.228
Observations	45895	45896	22641	22769

This table reports results of regression of log wages on caste and individual characteristics, using data from all 18-60 year old respondents of the 2011 IHDS. Wage data is taken from the respondent's highest income occupation, with the 1st and 99th percentiles removed. The "Occ. is caste's trad. occ." variable indicates the occupations (if any) that are traditionally performed by the respondent's caste, as defined in section 3. The caste-occ. network data variable is equal to the jacknifed ratio of the number of members of the respondent's caste in an occupation to the total workers in the occupation. The "SC" variable indicates that the respondent's reported caste is part of the state-level list of scheduled castes.
All specifications include controls for state fixed effects, education, age, experience, rural/urban location, OBC/SC/ST status, religion, missing paternal occupation, and a dummy variable for individuals who do not associate with a caste.
Standard errors clustered at the PSU (village) level. $*p < 0.10, **p < 0.05, ***p < 0.01$

Table 3: Returns to Human Capital in Traditional Occupations

	Log wages in chosen occupation					
		Male			Female	
	(1)	(2)	(3)	(4)	(5)	(6)
Years education	0.074***		0.074***	0.027		0.023
	(0.006)		(0.006)	(0.018)		(0.017)
Years education×	-0.032***		-0.032***	0.007		0.010
Occ. is any trad. occ.	(0.010)		(0.010)	(0.020)		(0.019)
Experience	0.028***		0.028***	0.005*		0.003
	(0.002)		(0.002)	(0.003)		(0.002)
Experience×	-0.019***		-0.019***	0.001		0.002
Occ. is any trad. occ	(0.003)		(0.003)	(0.003)		(0.003)
Father's occ.		0.134**	0.117**		0.393***	0.351***
		(0.060)	(0.055)		(0.136)	(0.121)
Father's occ.×		-0.071	-0.046		-0.184	-0.140
Occ. is any trad. occ		(0.063)	(0.058)		(0.140)	(0.127)
Caste-occ. network		1.949*	0.987		2.209***	2.139**
		(0.984)	(0.860)		(0.805)	(0.820)
Caste-occ. network×		-1.074	-0.115		0.548	0.561
Occ. is any trad. occ		(0.940)	(0.819)		(0.882)	(0.915)
Jati FE	Yes	Yes	Yes	Yes	Yes	Yes
Occupation FE	Yes	Yes	Yes	Yes	Yes	Yes
R-sq	0.292	0.273	0.293	0.268	0.268	0.273
Observations	45895	45895	45895	22641	22641	22641

This table reports results of regression of log wages on caste and individual characteristics, using data from all 18-60 year old respondents of the 2011 IHDS. Wage data is taken from the respondent's highest income occupation, with the 1st and 99th percentiles windsorized. The Occ. is any trad. occ" variable indicates whether an occuption is traditional for *any* caste, as defined in section 3. The caste-occ. network data variable is equal to the jacknifed ratio of the number of members of the respondent's caste in an occupation to the total workers in the occupation.
All specifications include controls for state fixed effects, education, age, OBC/SC/ST status, religion, missing paternal occupation, urban/rural location, and a dummy variable for individuals who do not associate with a caste. Standard errors are clustered at the PSU level. $*p < 0.10, * * p < 0.05, * * *p < 0.01$

Table 4: Wage Discrimination

	Log wages in chosen occupation			
	(1)	(2)	(3)	(4)
Female	-0.593***	-0.379***	-0.389***	-0.386***
	(0.027)	(0.018)	(0.018)	(0.018)
Other backwards caste	-0.057**	-0.066**	-0.049*	-0.051**
(OBC)	(0.024)	(0.027)	(0.025)	(0.025)
Scheduled caste	-0.055	-0.176***	-0.164***	-0.156***
(SC)	(0.037)	(0.041)	(0.039)	(0.037)
Scheduled tribe	-0.167***	-0.234***	-0.217***	-0.217***
(ST)	(0.040)	(0.044)	(0.043)	(0.042)
Father's occ.			0.105***	0.091***
			(0.014)	(0.014)
Caste-occ. network			1.295**	1.025**
			(0.510)	(0.455)
Occ. is caste's trad. occ.				0.123***
				(0.045)
Occupation FE	No	Yes	Yes	Yes
R-sq	0.224	0.307	0.309	0.310
Observations	68665	68665	68665	68665

This table reports results of regression of log wages on caste and individual characteristics, using data from all 18-60 year old respondents of the 2011 IHDS. Wage data is windsorized at the 1st and 99th percentile of the non-negative values within occupation category. The "Occ. is caste's trad. occ." variable indicates the occupations (if any) that are traditionally performed by the respondent's caste, as defined in section 3. The caste-occ. network data variable is equal to the jacknifed ratio of the number of members of the respondent's caste in an occupation to the total workers in the occupation.

All specifications include controls for state fixed effects, education, age, experience, religion, urban/rural location, missing paternal occupation, and a dummy variable for individuals who do not associate with a caste. Standard errors are clustered at the PSU level. $*p < 0.10, **p < 0.05, ***p < 0.01$

Table 5: Structural Parameters: Coefficients

Non-pecuniary utility (τ_{io})		General human capital $\left(\tilde{\beta}_i\right)$		Costs of education (κ_i)	
(1)		(2)		(3)	
Traditional occupation	0.178	Experience	0.111	Constant	-7.253
	(0.026)		(0.006)		(0.237)
Traditional occupation×	-0.086	Experience^2	-0.002	Females	1.347
female	(0.028)		(0.000)		(0.092)
Traditional occupation×	0.078	Education	0.336	Other backwards	0.119
OBC	(0.035)		(0.013)	caste	(0.170)
Traditional occupation×	-0.020	Experience×	-0.171	Scheduled caste	1.215
SC	(0.045)	female	(0.014)		(0.157)
Traditional occupation×	-0.133	Experience^2×	0.003	Scheduled tribe	1.465
ST	(0.052)	female	(0.000)		(0.187)
Homework×	3.991	Education×	-0.033	Constant×	0.583
female	(0.071)	female	(0.008)	education	(0.026)
Homework×	-0.137			Females×	0.071
female×OBC	(0.021)			education	(0.017)
Homework×	-0.381			OBC×	0.045
female×SC	(0.031)			education	(0.022)
Homework×	-0.726			SC×	-0.044
female×ST	(0.065)			education	(0.015)
				ST×	-0.007
				education	(0.024)

Occupation-specific human capital $\left(\tilde{\psi}_{io}\right)$		Labor force discrimination $(1 - T_{io})$	
(4)		(5)	
Father's occupation	1.462	Female	-1.102
×male	(0.025)		(0.052)
Caste's share in	8.156	OBC	0.052
occupation×male	(0.236)		(0.016)
Father's occupation	0.485	SC	0.027
×female	(0.045)		(0.016)
Caste's share in	1.280	ST	-0.050
occupation×female	(0.337)		(0.032)

Parameters displayed as estimated using the maximum likelihood estimator described in Section 6. Standard errors in parentheses clustered at the PSU level.

Table 6: Structural Parameters: Variances

	Parameter value
Occupational wage shocks (σ_π)	1.169
	(0.012)
Occupational wage shocks (σ_π)×	0.103
female	(0.013)
General skills (σ_α)×	1.890
female	(0.103)
Cost of education shocks (σ_κ)	1.812
	(0.121)
Cost of education shocks (σ_κ)×	1.593
female	(0.125)

Parameters displayed as estimated using the maximum likelihood estimator described in Section 6. Standard errors in parentheses clustered at the PSU level.

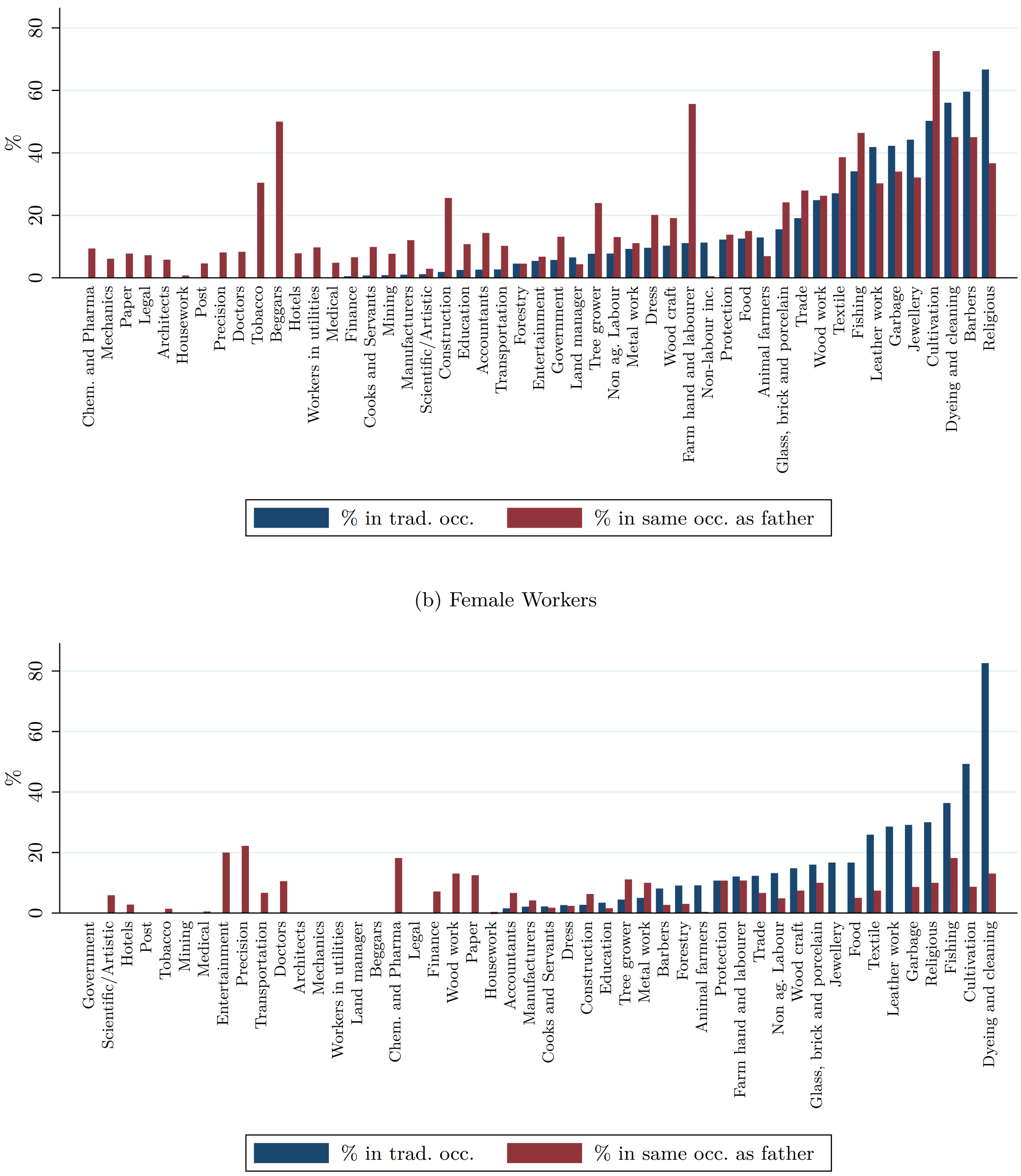
Table 7: **E**ffects of Removing Occupational Identity

		No occupational identity + endogenous wages (1)			(1) + endogenous education (2)			(2) + parental occupation orthogonal to trad. occ. (3)	
Panel A: Estimated Social networks									
i. Aggregate outcomes (percent changes from baseline)									
Market Output		0.062			0.346			-2.980	
Output per worker		0.335			0.629			-0.005	
Labor force participation		-0.273			-0.281			-2.975	
Schooling		0.000			0.489			1.048	
ii. Occupation/Caste-level Outcomes (percent changes)									
	Min	Median	Max	Min	Median	Max	Min	Median	Max
Occupation: wage rate	-0.092	-0.034	0.547	-0.275	-0.045	0.549	-1.602	-1.031	6.891
Occupation: human capital	-1.561	0.169	0.338	-1.289	0.498	1.178	-20.560	0.096	2.426
Occupation: employment share (pp)	-0.682	0.003	0.190	-0.718	0.004	0.195	-1.989	0.003	2.072
Occupation: trad. worker share (pp)	-5.195	-0.420	0.000	-5.349	-0.417	0.000	-12.410	-1.023	0.000
Caste: % workers in trad. occ. (pp)	-5.679	-0.243	0.335	-6.084	-0.246	0.342	-23.714	-0.790	0.759
Caste: total income	-0.884	-0.004	1.060	-0.785	0.074	3.381	-53.516	-0.420	13.538
Panel B: Endogenous Social Networks									
i. Aggregate outcomes (percent changes from baseline)									
Market Output		-0.145			0.760			-8.138	
Output per worker		0.168			1.056			-5.130	
Labor force participation		-0.313			-0.293			-3.171	
Schooling		0.000			0.667			1.892	
ii. Occupation/Caste-level Outcomes (percent changes)									
	Min	Median	Max	Min	Median	Max	Min	Median	Max
Occupation: wage rate	-0.259	-0.177	0.554	-0.973	0.004	0.732	-5.471	-3.085	11.077
Occupation: human capital	-1.787	0.394	0.636	-1.421	0.774	3.760	-32.971	1.027	8.754
Occupation: employment share (pp)	-0.631	0.004	0.239	-0.659	0.005	0.223	-2.044	0.005	2.422
Occupation: trad. worker share (pp)	-3.638	-0.298	0.000	-3.859	-0.299	0.000	-10.650	-0.699	0.000
Caste: % workers in trad. occ. (pp)	-5.933	-0.243	0.414	-6.170	-0.243	0.435	-27.482	-0.774	1.645
Caste: total income	-0.969	-0.110	0.997	-1.146	0.133	3.479	-56.123	-2.260	5.504

Results in Panel A are relative to a baseline economy simulated from the estimated parameter values (described in Section 6) and the same empirical social data used in the estimation. Results in Panel B are relative to the baseline with endogenous equilibrium caste networks for which we solve with all estimated parameters. All counterfactuals use posterior values values of α_i and values of η_i generated during estimation. All values are percent changes unless noted otherwise; pp denotes changes in percentage points.

10 Figures

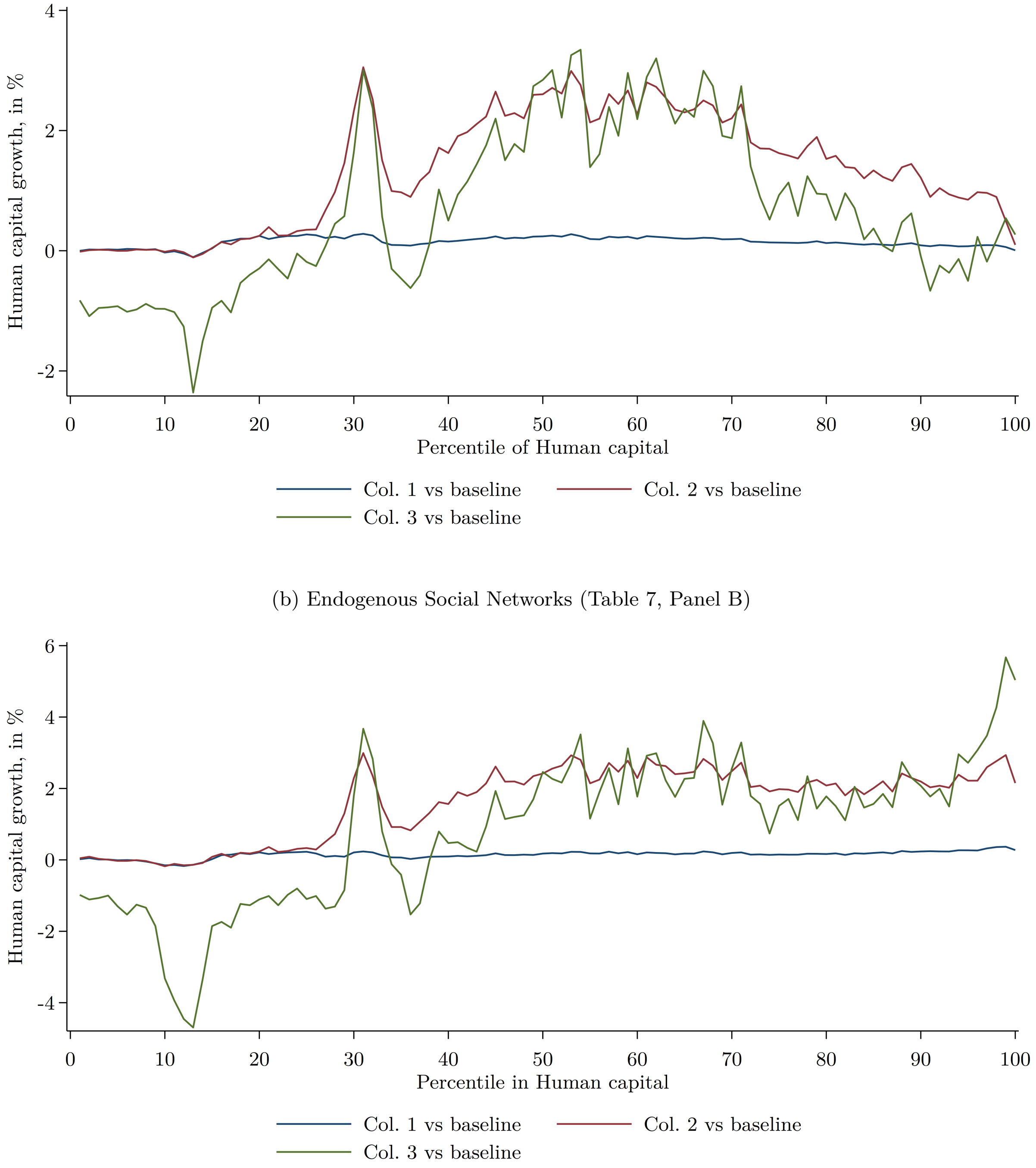
Figure 1: Occupational Composition: Traditional and Parental Transmission



(a) Male Workers

(b) Female Workers

Figure 2: Counterfactuals: Human Capital Growth Incidence Curves



(a) Exogenous Social Network (Table 7, Panel A)

(b) Endogenous Social Networks (Table 7, Panel B)

Appendix

A1 Data Appendix

The IHDS records household and individual income streams from a variety of different income sources. Because we are interested only in the active population, we restrict our sample to individuals aged 18 to 60 and drop full-time students and the unemployed.

While the survey provides the time spent and the income earned from most occupations at the individual level, for certain occupations, we had to make assumptions in order to derive an individual level wage.

First, the income derived from certain occupations, such as household businesses, are reported exclusively at the household level. Since we know how much time is spent in each occupation, but not the productivity of each household member in the household occupation, we attribute the same hourly income to each individual of the household working in the household occupation.

Second, the IHDS does not provide information on the time spent in animal care, but does provide the number of animals of each type owned by the household, as well as individual level information recording which household members take care of animals. In order to derive hourly information, we turn to an additional household survey, the REDS 2006 data. This survey data, representative at rural India level contains information on the time spent on animal care by household members, as well as on the number of animals of each type owned by a household. We predict the time spent by IHDS household member in animal care by using the coefficient of an OLS regression of the time spent on animal care of REDS household members on the number of each type of animal and their square.

Third, the IHDS does not provide the time spent in occupations such as money lending and the rental of land. Since we did not find alternative data sources that would allow to infer the time spent in these activities, we had to derive them based on the time spent and income earned by the same individual in other occupations. For individuals that earn income from other occupations and from which we know the time spent in these other occupations, we compute the time spent in money lending and the rental of land so that the share of time spent in these occupations equals the share of these occupations in the individual's total income (that is, we assume that the productivity in these activities in the same as the average productivity in the individual's other activities). We attribute income from these activities to the head of the household if he is 60 years old or younger, otherwise to his eldest son in the household.

Fourth, we find that, among individuals 18 to 60 years old, almost 34% report their "primary activity" as housework, with these respondents being almost entirely women. However, when further probed, many women reportedly engaged in housework indicate that they spend many hours working for income. We therefore designate a respondent's main occupation as housework only if she works less than the median number of hours in all occupations in which she reports working (or reports no other work). Otherwise we assign her to her highest income-earning occupation as described above. This results in a final sample in which 12.25% of respondents have housework as their occupation.

Fifth, because many individuals report multiple sources of income, we designate each individual's occupation as the activity from which they receive the most income and spend the most time. If these are different, we choose the occupation broadly identified in the survey as the "main activity" of the household, then the activity to which the respondent devotes the most time.

Sixth, income per hour are trimmed at the bottom 1% and top 99%. Results are robust, and in fact stronger, when trimming at the bottom 0.1% and top 99.9% instead.

A2 Additional empirical results

A2.1 Robustness of occupational choice and wage results

To confirm the robustness of our results on the importance of traditional occupations for the allocation of human capital, we carry out a variety of robustness checks. Table A1 presents the results of a robustness check on our occupational choice analysis presented in Table 1. We show that the occupational choice results remain robust to the inclusion of additional controls. We successively add controls for: whether the occupation is considered ritually polluting (impure) and the individual is member of a forward caste, whether the occupation is pure (i.e. a occupation traditionally attached to high castes) and the individual is a SC, whether the occupation is agricultural and the individual inherited land. Finally, to allow for intrafamilial transfer of human capital on top of direct transfer from father to child, we additionnally control for wether the occupation is the individual's uncle's occupation. None of these additional controls seems to alter the fact that individuals are more likely to work in their jati's traditional occupation.

A2.2

A3 Model Appendix

A3.1 Expected lifetime utility

In this section we derive expected lifetime utility. We follow Mincer's original work and assume that all individuals work for T years after finishing schooling. Lifetime utility at the time of the schooling choice is the discounted sum of utility starting immediately after schooling ends (i.e. $t = s$) until the end of the working period (i.e. $s + T$),

$$U_i^* \quad = \quad \max_s \left\{ \mathbb{E}_{\pi_{io}} \left[\max_o \left\{ \int_s^{s+T} e^{-rt} \left(\log \left((1 - T_k) \, w_o \psi_{io} \left(\alpha_i \beta_i \right)^{\rho_o} \pi_{io} \right) + \tau_{io} + A_o \right) dt \right\} \right] \right\}$$

We express observable human capital, β_i, as a function of education s_i and a quadratic function of experience:

$$\beta_i = \exp \left(\tilde{\beta}_s s_i + \tilde{\beta}_x^1 \left(t - s_i - b \right)^1 + \tilde{\beta}_x^2 \left(t - s_i - b \right)^2 \right),$$

where $(t - s_i - b)$ is individuals' experience equal to individuals' age t minus their years of schooling s_i and minus the age at which individuals typically begin school, which we denote by b and which we assume to be equal to 6. The $\tilde{\beta}_s$ and $\tilde{\beta}_x$ coefficients are parameters that map years of schooling and experience into human capital units.

Integrating over years of expected labor force participation yields,

$$\begin{aligned} U_i^* &= \bar{r}\mathbb{E}_{\pi_{io}}\left[\max_o \left\{\log\left(\left(1-T_k\right) w_o \psi_{io}\left(\alpha_i \bar{\beta}_i\right)^{\rho_o} \pi_{io}\right) + \tau_{io} + A_o\right\}\right] \\ &\equiv \bar{r}\mathbb{E}_{\pi_{io}}\left[\max_o \left\{\bar{u}_{io} + \log\left(\pi_{io}\right)\right\}\right], \end{aligned}$$

where $\bar{r} = \frac{e^{-rs}}{r}\left(1-e^{-rT}\right)$ is the discount factor that incorporates years spent in school, and $\bar{\beta}_i = \exp\left(\tilde{\beta}_s s_i + \bar{\beta}_x\right)$ where $\bar{\beta}_x$ is pre-employment expected value of experience and equal to:

$$\bar{\beta}_x = -\tilde{\beta}_x^1 b + \tilde{\beta}_x^2 b^2 + \left(\left(1-e^{-rT}-e^{-rT} rT\right)\left(\tilde{\beta}_x^1 r + \tilde{\beta}_x^2\left(-r2b+2\right)\right) - e^{-rT} r^2 T^2 \tilde{\beta}_x^2\right) / \left(1-e^{-rT}\right) r^2.$$

A3.2 Educational Choice

Children choose their years of schooling s_i to maximize discounted lifetime utility net of schooling costs. We assume that occupation-specific skill shocks π_{io} are not known at this time, so that individuals form expectations about their future occupational choice probabilities based on their knowledge of their other characteristics, their caste affiliations and their parental occupation. Children therefore solve:

$$\begin{aligned} V_i^* &= \max_s \left\{\bar{r}\mathbb{E}_{\pi_{io}}\left[\max_o \left\{\bar{u}_{io} + \log\left(\pi_{io}\right)\right\}\right] - \left(\kappa_{1k} + \frac{\kappa_{2k}}{2} s_i + \eta_i\right) s_i\right\} \\ &= \max_s \left\{\frac{\bar{r}}{\sigma_\pi} \log \sum_o \exp\left(\sigma_\pi \bar{u}_{io}\right) - \left(\kappa_{1k} + \frac{\kappa_{2k}}{2} s_i + \eta_i\right) s_i\right\}, \end{aligned}$$

which yields the following first order condition:

$$\left(\kappa_{1k} + \kappa_{2k} s_i + \eta_i\right) + \frac{\bar{r}}{\sigma_\pi} r \log\left[\sum_o \left(\sigma_\pi \bar{u}_{io}\right)\right] = \bar{r}\tilde{\beta}_s \sum_o \rho_o P_{io}.$$

Individuals choose their level of schooling to equate marginal costs of schooling (LHS) with marginal returns (RHS). Schooling costs depend on the direct costs (κ and η) and the opportunity cost from foregone income. Returns to schooling are given by the returns to schooling $\tilde{\beta}_s$ multiplied by the probability weighted occupation-specific returns to human capital ρ_o.

A3.3 Equilibrium

We first summarize all exogenous model parameters before defining the equilibrium. The parameters $\{\tilde{\beta}, \tilde{\psi}, \rho_o\}$ matter for determining worker i's productivity in each occupation.[28] In addition, entrepreneurs'

[28] Recall that β_i captures general human capital of a worker from observable characteristics, ψ_{oi} captures productivity shifters from caste-occupation networks and father's occupation, ρ_o captures the occupation-specific returns to general human capital. We discuss the specific parameterizations that link these variables to the data in Section 6.

disutility from hiring certain castes in certain occupations δ_{ok} generates wage discrimination and therefore affects castes' effective occupational wage rate per human capital unit. Exogenous parameters from the production function are total factor productivity A, occupational factor shares Z_o and the elasticity of substitution between occupations σ. A worker's utility depends on occupation amenities A_o and his preference for working in his castes' traditional occupation τ_{ok}. Last, the parameters σ_π and σ_α respectively characterize the dispersion of the idiosyncratic productivity shocks π_{io} and α_i. We denote the full set of exogenous parameters by: $\Omega = \left\{\tilde{\beta}, \tilde{\psi}, \rho_o, \delta_{ok}, A, Z_o, \sigma, A_o, \tau_{ok}, \sigma_\pi, \sigma_\alpha\right\}$.

Given exogenous parameters Ω, the equilibrium of the economy is characterized by:

1. Occupational choice probabilities P_{io} that are consistent with individuals' utility maximization as derived in Equation 5.

2. Education choices s_i that are consistent individuals' utility maximization shown in Equation 7.

3. Firms' hire human capital Θ_o in each occupation that is consistent with their profit maximization from Equation 10.

4. Due to perfect competition, wage discrimination exactly offsets entrepreneurs' disutility of hiring certain castes, so that $T_k = \delta_k$ which makes entrepreneurs indifferent between hiring workers from all castes.

5. Wage rates per human capital unit w_o clear occupational labor markets, ensuring that human capital demand (derived from entrepreneurs hiring choices) equals human capital supply (derived from individual choices in Equation 9) in each occupation.

6. Good market clears so that total consumption equals total output (cf. aggregate CES production function shown in Equation 10).

A3.4 Derivation of the wage distribution and likelihood

We derive the formulas for the likelihood function of the observed occupation, wages and education. We proceed in two steps: first, we derive the distribution of occupation-specific skill shocks conditional on having chosen an occupation. Second, we use this to derive the distribution of wages, again conditional on having chosen an occupation.

Let V_i^* be the maximum utility at the moment of the occupational choice for an individual that chooses occupation o^* to maximize his utility:

$$V_i^* = \max_o [V_{io}] = \max_o [\bar{u}_{io} + \log(\pi_{io})] = \bar{u}_{io}^* + \log(\pi_{io}^*).$$

Under Assumption 1, $\log(\pi_{io})$ is Gumbel distributed, which implies that the maximum utility level

V_i^* is also Gumbel distributed since:

$$\begin{aligned}\Pr(V_i^* \leq x) =& \Pr\left(\bar{u}_{io} + \log(\pi_{io}) \leq x\right) \forall o \\ =& \prod_{o'} \exp\left\{-\exp\left(-\sigma_\pi (x - \bar{u}_{io'})\right)\right\} \\ =& \exp\left\{-\exp\left(-\sigma_\pi \left[x - \frac{1}{\sigma_\pi}\log\sum_{o'}\exp\left(\sigma_\pi \bar{u}_{io'}\right)\right]\right)\right\},\end{aligned}$$

which corresponds to the Gumbel CDF with location $\frac{1}{\sigma_\pi}\log\left(\sum_{o'}\exp\left(\sigma_\pi \bar{u}_{io'}\right)\right)$ and shape parameter σ_π. Using this result, we can now derive the distribution of occupation-specific skill shocks π_{io}^* conditional on having chosen occupation o in the following way:

$$\begin{aligned}H_i(x) = \Pr(\pi_{io}^* \leq x | V_{io} = V_i^*) =& \Pr\left(\frac{\exp(V_i^*)}{\exp(\bar{u}_{io}^*)} \leq x\right) \\ =& \exp\left\{-\exp\left(-\sigma_\pi \log\left[x \exp\left(\bar{u}_{io}^*\right)\right] + \log\left[\sum_{o'}\exp\left(\sigma_\pi \bar{u}_{io'}^*\right)\right]\right)\right\} \\ =& \exp\left\{-x^{-\sigma_\pi}\left(P_{io}^*\right)^{-1}\right\}.\end{aligned}$$

This shows that skill shocks π_{io}^* are Frechet distributed with the mean being the inverse of the occupational choice probability $\left(P_{io}^*\right)^{-1}$ (cf. Equation 5). Using this result, we now derive the distribution of observed earnings $\hat{y}_{io}$ conditional on individuals' occupational choice–recalling also that $y_{io} = (1 - T_k)\, w_o \left(\alpha_i \beta_i\right)^{\rho_o} \psi_{io} \pi_{io}$, so that:

$$\begin{aligned}J_i(x) &= \Pr\left(y_{io} \leq x | V_{io} = V_i^*\right) = \Pr\left(y_{io}^* \leq x\right) = \Pr\left((1 - T_k)\, w_o \left(\alpha_i \beta_i\right)^{\rho_o} \psi_{io} \pi_{io}^* \leq x\right) \\ &= \exp\left(-\left(\frac{x}{(1 - T_k)\, w_o \left(\alpha_i \beta_i\right)^{\rho_o} \psi_{io}}\right)^{-\sigma_\pi}\left(P_{io}^*\right)^{-1}\right) \\ &= \exp\left(\frac{-\sum_o \left(\exp(\bar{u}_{io'}^*)\right)^{\sigma_\pi}}{\left(\exp\left(\tau_{io} + A_o + \rho_o\left(\bar{\beta} - \beta_i\right)\right) x\right)^{\sigma_\pi}}\right).\end{aligned}$$

Last, we take the derivative to obtain the PDF of observed income:

$$\begin{aligned}\Pr(y_{io}^* = x | V_{io} = V_i^*) =& \frac{d}{dx} J_i(x) \\ =& \frac{\sigma_\pi}{x} \frac{\sum_{o'}\left(\exp(\bar{u}_{io'}^*)\right)^{\sigma_\pi}}{\left(\exp\left(\tau_{io} + A_o + \rho_o\left(\bar{\beta} - \beta_i\right)\right) x\right)^{\sigma_\pi}} \exp\left(\frac{-\sum_{o'}\left(\exp(\bar{u}_{io'}^*)\right)^{\sigma_\pi}}{\left(\exp\left(\tau_{io} + A_o + \rho_o\left(\bar{\beta} - \beta_i\right)\right) x\right)^{\sigma_\pi}}\right).\end{aligned}$$

A4 Additional counterfactual results

A4.1 Algorithm

We implement the counterfactual analysis through a fixed point algorithm. We first modify parameters or variables according to each counterfactual scenario. With exogenous (fixed) caste-occupation networks, we simply iterate on the human capital distribution across occupations–and hence the occupational wage rates implied by market clearing–until they are consistent with individuals' optimal education and occupational choices. With endogenous caste-occupation networks, we add a second fixed point where we update caste-occupation networks based on individuals' occupational choices in an outer loop. Hence, we iterate on caste-occupation networks and the human capital distribution across occupations.

A4.2 Counterfactual results with weaker network effects

This sensitivity analysis also addresses the potential concern that estimates of social network effects on productivity may be biased. Unobservable characteristics could for example make certain castes more productive in certain occupations–attracting more caste members to these occupations and increasing their wages. We explore the potential biases this might cause by recomputing all counterfactuals with the strength of network effects reduced by half, formally dividing $\tilde{\psi}_2$ and $\tilde{\psi}_4$ by 2. Table A3 shows that our results with weaker network effects are roughly similar in magnitude.

A5 Appendix Tables

Table A1: Robustness: Occupational Choice

	Probability of occupational choice- male (N =2,269,092)					
	(1)	(2)	(3)	(4)	(5)	(6)
Occ. is caste's trad. occ.	0.047***	0.047***	0.047***	0.045***	0.046***	0.045***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Occ. is father's occ.	0.306***	0.306***	0.306***	0.302***	0.306***	0.301***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Occ. is caste's trad. occ.×SC	-0.026***	-0.026***	-0.026***	-0.024***	-0.026***	-0.024***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Impure occ.×FC		-0.000				0.000
		(0.000)				(0.000)
Pure occ.×SC			-0.002***			-0.002***
			(0.001)			(0.001)
Agricultural occ.× land inherited				0.007***		0.007***
				(0.001)		(0.001)
Occ. is uncle's occ.					0.150***	0.136***
					(0.022)	(0.021)
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes
Occ. FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.154	0.154	0.154	0.157	0.155	0.157

This table reports results of a linear probability model of occupational choice, using data from all 18-60 year old respondents of the 2011 IHDS. The dataset contains all unique combinations of respondent and occupation, with the outcome variable equal to 1 for the occupation chosen by the respondent, and 0 for all other occupations. The "Occ. is caste's trad. occ." variable indicates the occupations (if any) that are traditionally performed by the respondent's caste, as defined in section 3. The caste-occ. network data variable is equal to the jacknifed ratio of the number of members of the respondent's caste in an occupation to the total workers in the occupation. The "SC" variable indicates that the respondent's reported caste is part of the state-level list of scheduled castes.
Standard errors clustered at the PSU (village) level. $*p < 0.10, **p < 0.05, ***p < 0.01$

Table A2: Occupation-level Structural Parameters

	Occupation skill-wage ($\ln w_o$)	Occupation Amenity ($\ln A_o$)	Returns to skill ρ_o
	(1)	(2)	(3)
Non-labour income earner	-3.705 (0.117)	3.713 (0.091)	0.374 (0.032)
Cultivation	-0.573 (0.052)	3.365 (0.063)	-0.035 (0.017)
Land manager	-3.936 (0.392)	2.658 (0.261)	0.127 (0.069)
Farm hand and labourer	0.874 (0.035)	2.087 (0.068)	-0.210 (0.011)
Animal farmers	-0.858 (0.045)	3.946 (0.055)	-0.517 (0.036)
Tree and Shrub Crop Growers	-1.510 (0.184)	1.932 (0.089)	-0.220 (0.033)
Fish related workers	-2.188 (0.267)	2.129 (0.142)	-0.071 (0.056)
Forest hunters, gatherers and officers	-1.891 (0.259)	1.982 (0.205)	0.033 (0.040)
Mining related worker	-2.590 (0.235)	2.023 (0.128)	0.119 (0.048)
Labourers, non-agricultural	-0.508 (0.067)	2.063 (0.069)	0.064 (0.015)
Chemical and pharma related worker	-3.770 (0.544)	2.163 (0.124)	0.306 (0.110)
Textile related worker	-1.443 (0.136)	2.224 (0.087)	0.039 (0.029)
Wooden crafts and instruments	-2.448 (0.176)	2.373 (0.151)	-0.150 (0.059)
Dyeing, cleaning and washing related worker	-3.835 (0.281)	2.227 (0.169)	0.108 (0.070)
Dress related workers	-0.811 (0.136)	2.208 (0.079)	0.007 (0.032)
Leather workers	-3.547 (0.321)	2.314 (0.151)	0.228 (0.053)
Wood items related worker	-1.321 (0.101)	1.986 (0.083)	0.186 (0.018)
Metal related worker	-1.753 (0.107)	2.188 (0.068)	0.277 (0.020)
Glass, brick and porcelain related worker	-2.082 (0.169)	2.084 (0.087)	-0.089 (0.034)
Food and beverage producers	-2.043 (0.154)	2.345 (0.112)	0.092 (0.035)
Tobacco products	-1.899 (0.137)	2.947 (0.083)	-0.525 (0.030)
Barbers and beauticians	-2.770 (0.155)	2.332 (0.104)	0.228 (0.030)
Construction	0.951 (0.040)	2.044 (0.069)	0.018 (0.008)
Workers in utilities (power, water, etc)	-2.692 (0.124)	2.271 (0.085)	0.515 (0.025)
Printers, paper and book makers	-4.500 (0.411)	2.548 (0.132)	0.531 (0.070)
Precision Instrument Makers and Repairers	-2.725 (0.133)	2.408 (0.102)	0.445 (0.030)
Jewelers and Precision Metal Workers	-2.903 (0.225)	2.031 (0.111)	0.321 (0.035)
Garbage workers	-1.036 (0.118)	1.795 (0.098)	-0.080 (0.028)
Transportation of all kinds	-0.841 (0.081)	2.272 (0.078)	0.288 (0.015)
Post office, Telegraph and Telephone service	-5.054 (0.362)	2.314 (0.147)	0.710 (0.058)
Financial intermediation	-6.340 (0.267)	3.675 (0.183)	0.874 (0.047)
Trade and retail shops	-1.114 (0.092)	2.670 (0.083)	0.354 (0.022)
Hotels	-3.022 (0.219)	2.456 (0.109)	0.228 (0.057)
Music and entertainment	-4.022 (0.504)	2.319 (0.263)	0.364 (0.090)
Protective services	-3.128 (0.118)	2.410 (0.090)	0.525 (0.029)

Table A2: Occupation-level Structural Parameters

	Occupation skill-wage ($\ln w_o$)	Occupation Amenity ($\ln A_o$)	Returns to skill ρ_o
	(1)	(2)	(3)
Government service	-6.112 (0.514)	2.683 (0.204)	0.880 (0.073)
Religious workers	-5.093 (0.332)	2.681 (0.175)	0.469 (0.061)
Legal professionals	-9.213 (0.417)	3.476 (0.165)	1.136 (0.055)
Doctors, modern and traditional	-8.412 (0.584)	3.151 (0.247)	1.090 (0.072)
Other medical professionals	-7.336 (0.416)	3.214 (0.154)	1.129 (0.060)
Professors, teachers, education professionals	-7.755 (0.394)	3.674 (0.135)	1.285 (0.066)
Accountants, secretaries, clerks	-3.589 (0.196)	2.633 (0.111)	0.821 (0.035)
Architects, surveyors, engineers, and their employees.	-5.555 (0.187)	2.102 (0.112)	0.949 (0.039)
High skill scientific or artistic	-4.723 (0.304)	2.564 (0.112)	0.633 (0.055)
Cooks and house servants	-0.417 (0.060)	2.232 (0.075)	-0.283 (0.022)
Manufacturers, business men and contractors otherwise unspecified	-2.858 (0.186)	2.469 (0.102)	0.516 (0.033)
Mechanics otherwise unspecified	-3.996 (0.311)	2.495 (0.144)	0.434 (0.061)
Home work	0 (normalized)	0 (normalized)	0.327 (0.014)

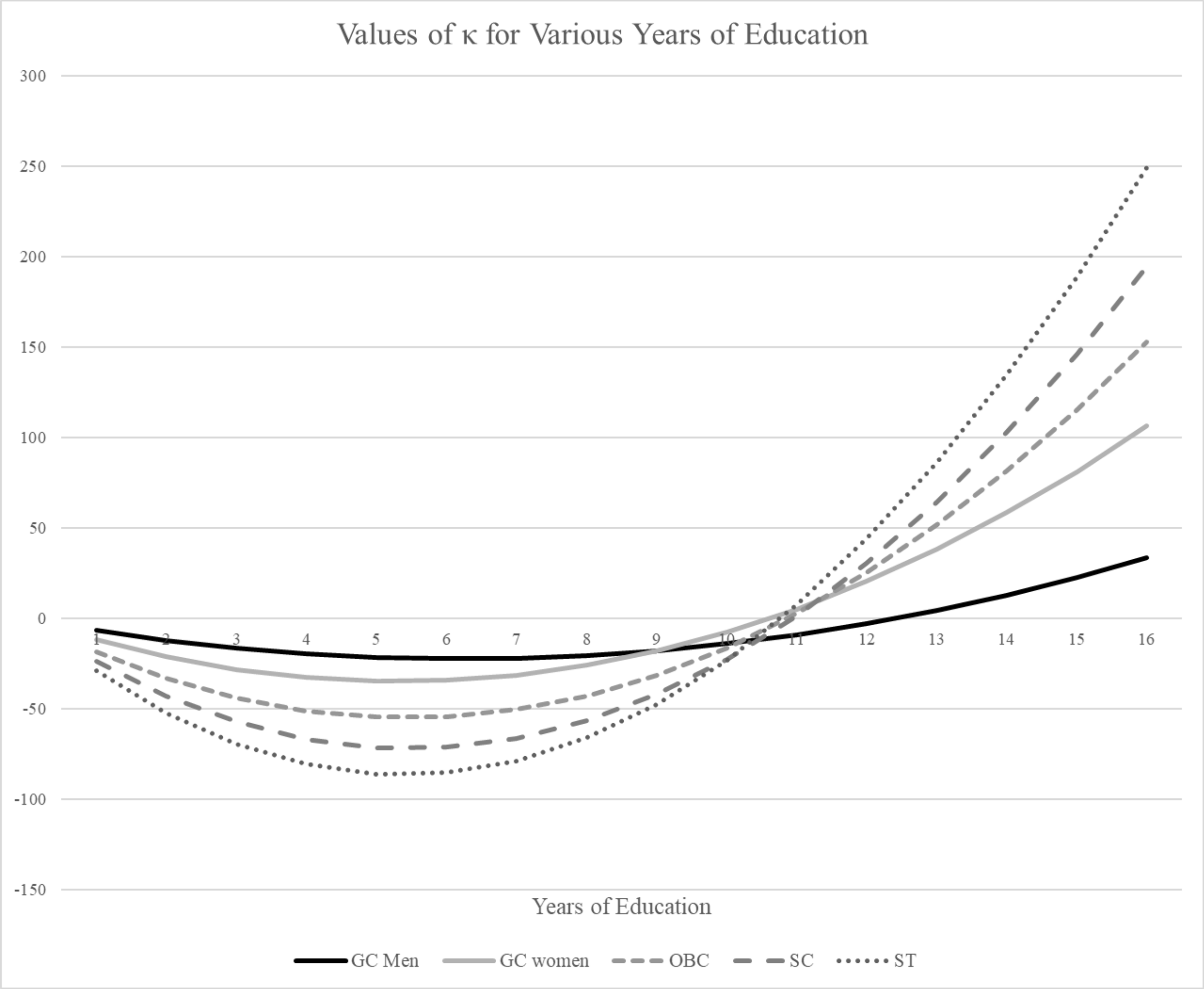
Table A3: **Effects of Removing Occupational Identity: Weaker network effects**

		No occupational identity + endogenous wages (1)			(1) + endogenous education (2)			(2) + parental occupation orthogonal to trad. occ. (3)	
Panel A: Estimated Social networks									
i. Aggregate outcomes (percent changes from baseline)									
Market Output		0.064			0.332			-2.764	
Output per worker		0.318			0.595			0.081	
Labor force participation		-0.254			-0.261			-2.842	
Schooling		0.000			0.463			0.977	
ii. Occupation/Caste-level Outcomes (percent changes)									
	Min	Median	Max	Min	Median	Max	Min	Median	Max
Occupation: wage rate	-0.092	-0.028	0.298	-0.263	-0.038	0.420	-1.507	-0.949	6.303
Occupation: human capital	-0.824	0.155	0.340	-0.920	0.466	1.127	-19.055	0.086	2.301
Occupation: employment share (pp)	-0.672	0.003	0.175	-0.705	0.004	0.180	-1.916	0.002	1.961
Occupation: trad. worker share (pp)	-4.286	-0.393	0.000	-4.391	-0.396	0.000	-9.456	-0.929	0.000
Caste: % workers in trad. occ. (pp)	-5.677	-0.241	0.335	-6.084	-0.240	0.341	-23.692	-0.742	0.755
Caste: total income	-0.879	-0.003	1.061	-0.785	0.071	3.382	-53.494	-0.385	13.602
Panel B: Endogenous Social Networks									
i. Aggregate outcomes (percent changes from baseline)									
Market Output		0.006			0.233			-2.855	
Output per worker		0.262			0.501			-0.080	
Labor force participation		-0.256			-0.267			-2.777	
Schooling		0.000			0.459			0.949	
ii. Occupation/Caste-level Outcomes (percent changes)									
	Min	Median	Max	Min	Median	Max	Min	Median	Max
Occupation: wage rate	-0.123	-0.062	0.960	-0.274	-0.068	0.806	-1.486	-1.032	6.203
Occupation: human capital	-2.821	0.195	0.376	-2.153	0.465	1.061	-18.903	0.337	2.177
Occupation: employment share (pp)	-0.671	0.003	0.174	-0.705	0.004	0.182	-1.866	0.002	1.895
Occupation: trad. worker share (pp)	-4.631	-0.396	0.000	-4.700	-0.392	0.000	-9.929	-0.880	0.000
Caste: % workers in trad. occ. (pp)	-5.685	-0.230	0.334	-6.093	-0.230	0.339	-23.661	-0.682	0.747
Caste: total income	-0.900	-0.009	1.051	-1.311	0.053	3.407	-53.317	-0.396	13.565

The counterfactuals shown here are computed in the same way as in Table 7, except that we now reduce the strength of caste-occupation networks by half, formally dividing $\tilde{\psi}_2$ and $\tilde{\psi}_4$ by 2.

A6 Appendix Figures

Figure A1: Values of κ


Values of κ for Various Years of Education
Years of Education
GC Men
GC women
OBC
SC
ST