A Division of Laborers: Identity and Efficiency in India

Guilhem Cassan
University of Namur

Daniel Keniston
Louisiana State University

Tatjana Kleineberg
World Bank*

January 8, 2021

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 human capital allocation 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 which enable 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, SEA and ASSA 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 caste and occupation has often been predicted (Srinivas, 2003), it still 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 human capital across occupations. Second, the sorting of castes into traditional occupations can enable the transfer of occupational knowledge and human capital from parents to their children and the formation of social networks in traditional occupations. These two channels shape the economy in ways that can lastingly affect workers' occupational choices and can create "path dependence" well beyond workers' own preferences: 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 of working in the same occupation as their fathers. Unlike the distortionary effects on the allocation of human capital, 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 and wages. 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 work in other occupations. However, when we examine earnings within occupation (i.e., controlling for occupation 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 are informative about how workers select into occupations based on their comparative and absolute advantage, which we formalize in our model.

We develop and estimate a structural general equilibrium Roy (1951) model of education and occupational choice that incorporates caste identity through several channels: a direct preference for traditional occupations, productivity effects from working in one's father's occupation, and network effects at the caste-occupation level. As in the standard Roy model, workers differ in productivity which varies independently across occupations so that workers select into occupations based on their comparative advantage. The marginal workers who select into their traditional occupation because of their caste identity are therefore less productive than workers who choose the same occupation solely based on comparative advantage. We extend the model by allowing workers to also differ in general ability. This can induce the opposite type of selection because traditional occupations have, as we show, lower returns to ability. Workers with high general ability therefore prefer to work in "modern" occupations where returns to ability are higher, but their caste identity may draw them back into their traditional low-return occupation. Allowing workers to differ in occupation-specific and general ability can reconcile our empirical findings described above.

To study the importance of castes' affinity for their traditional occupations 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, it is essential to allow wages to adjust when estimating the effects of reallocating human capital across occupations. In addition, occupational choices are closely linked to individuals' education choices and to the composition of social networks. We therefore determine wages, educational choices, 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, that can help 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 historically seen as ritually superior and others 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 face wage discrimination and different costs of acquiring education.

All of these channels affect individuals' education and occupational choices, which jointly 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 we use historical sources to link this information to detailed data on castes' traditional occupation. We then use our estimated model to investigate the

importance of occupational caste identity for occupational distributions, wages, aggregate output and aggregate productivity. We do so by successively removing channels that link castes to their traditional occupations. Specifically, we remove castes' attachment to traditional occupations, first holding caste networks constant, and then endogenizing networks as a function of workers' occupational choices.

We find very small aggregate effects with constant and endogenous networks: output per worker increases respectively by 0.6 and 1.1 percent and aggregate output by 0.3 and 0.8 percent. Effects are small because improvements in workers' selection based on their comparative advantage are offset by reduced productivity from weaker caste networks and less intergenerational knowledge transfer since fewer workers work in their fathers' occupation. Despite the trivial aggregate effects, we nevertheless find large distributional effects: The share of traditional workers decreases by roughly 10 percent in the aggregate, by 4-5 percentage points in the most affected occupation and by roughly 6 percentage points in the most affected caste.

We then additionally eliminate the last link between castes and their traditional occupations by removing the correlation between fathers' occupations and traditional occupations. With constant caste networks, this leads to a substantial drop in output (-3 percent) and a small drop in output per worker (-0.01 percent). Effects are negative because we remove the option of simultaneously choosing one's father's and one's traditional occupation–where networks are largest. With endogenous caste networks, losses are even larger: total output drops by 8.1 percent and output per worker by 5.1 percent. Caste networks are now weaker, since we have removed all coordinating elements that previously clustered caste networks in their traditional occupations, which lowers productivity and output. These findings indicate that reduced productivity through weaker networks and less intergenerational learning can dominate over gains from an improved 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 which study 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), national identity and support for redistribution (Shayo, 2009), as well as attachment to hometown and reduced occupational and geographical mobility (Munshi and Wilson, 2011). Another relevant literature examines the intergenerational transmission of education and occupation choices. Altonji and Dunn (2000) and 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. However, several theoretical studies have investigated 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 in the labor market if such choices imply deviating from social norms. In line with our findings, Akerlof (1976) points out that removing a taste for following widely shared social norms may not be enough to alter behavior 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 that 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 young men from pursuing high skill occupations in order to preserve strong social networks in low skill traditional occupations. A recent comprehensive survey of this literature 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 show that intergenerational transmission is remarkably strong for 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). Kumar et al. (2002) and Vaid (2012) document that castes play an important role in the intergenerational transmission of occupation, which does not seem to weaken over time.

Our work further 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 et al. (2019), which quantifies the effect of decreased discrimination against women and blacks in high skill-return US occupations on aggregate wage and GDP growth. It is common in this literature to assume that occupation-specific productivity is uncorrelated across occupations. Under this assumption, work-

[1] See also the literature on the pattern of caste based patron-client relationships and occupational specialization, the "jajmani" system, initiated by Wiser (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. In line with our results, this study demonstrates that traditionally agricultural castes perform better in agriculture than others, and are less likely to enter off-farm occupations.

ers' 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 model specifications for occupational choices and wage distributions across and within castes. One contribution of our paper is the combination of a general equilibrium approach with a structural estimation by maximum likelihood that relies on rich individual-level data.

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 several castes. Datasets that contain information on 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 with data 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:

The IHDS and DHS both 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 which was launched in 1985 by the Anthropological Survey of India and which made an extraordinary effort to systematically collect data on all Indian jatis. The project produced a volume (Singh, 1996) that lists all jati names and their various synonyms at the state level. We digitized this volume to create

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

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 an extensive occupation and income module that records income and time spent in each occupation for each individual in the household.[4] The survey also documents the occupation of the household head's father.[5] When father's occupation is missing, we impute it by constructing occupational probabilities based on the data of fathers' occupations from members of the same caste.[6] In Appendix A2.2, we use additional data from a different survey, which provides full data on all parents' occupations, to show that our main results are not affected by this imputation. 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. To do so, we follow Cassan (2019) and use the classifications from official state-wise lists of reservation.

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) which provides caste and occupation information for a very large sample of 109,041 households. We use the DHS only to supplement the social network data because it contains very sparse information on income and parental occupation, so that we cannot use it for the main part of our analysis.

Traditional occupations from historical data sources:

Our main source of information on jatis' 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 when necessary to complement data that is missing or incomplete data in the second round. In particular, we use jati names from the first round if they are missing or coded in a very general way, such as "scheduled caste" (SC), in the second round. Similarly, we use parental occupation from the first round if it is missing in the second round. 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 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 in the IHDS and DHS datasets, we create 48 consistent occupational categories (see list in Appendix Table A3).

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 women. 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. Overall, 16.8 percent of men work in their jati's traditional occupation, versus 9.1 percent if workers were randomly allocated across occupations (keeping the occupational distribution constant).[8] For women, 8 percent are in their jati's traditional occupation against 5 percent if randomly allocated.[9] Workers also tend 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, but these mechanisms are not perfectly correlated.

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, such as fishing, jewelry, or cultivation. Oh (2019) confirms 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

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 20 and 13 percent for rural ares and 13 and 4 percent for urban areas.

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

individual is observed once for each potential occupation (i.e., 48 times). We then run the following OLS regression:

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

where Occ_{iok} indicates that individual i of jati k works in occupation o, $TradOcc_{ok}$ 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 it 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 who belong 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: the probability that male workers choose an occupation increases by 31 percentage points if it is their father's occupation; and by 10 percentage points for each 1 percentage point increase in the occupation's caste network (Column 3). The impact of traditional occupations remains significant with these controls 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 workers of scheduled castes (SCs) have less affinity for their traditional occupations, although these results are imprecisely estimated (see Column 4).

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 0.6 percentage points when controlling for father's occupation and caste-occupation networks. Women's occupational choice probability increases by 14 percentage points for their father's occupation and by 3 percentage points for each 1 percentage point increase in the occupation's caste network (Column 3).

Selection and productivity in traditional occupations

To examine the relationship between occupational identity and hourly wages, 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 occupation o is the traditional occupation of worker i's jati, and X_{iok} is a set of individual characteristics which include father's occupation and caste networks. Table 2 presents the results where we consider two specifications, controlling first for jati fixed effects and then for occupation fixed effects.

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.[11] 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.[12]

Perhaps surprisingly, the results are reversed with occupation fixed effects (Columns 2 and 4). Here we find that workers in their traditional occupation earn 11-13 percent more per hour than non-traditional workers in the same occupation. This result is at odds with the selection 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.

Returns to ability in traditional occupations

To explain these findings, we examine differences in returns to ability as a potential driver of occupational choices and wages. We test whether traditional occupations–which existed by definition in pre-industrial times–have different returns to ability than "modern" occupations by running 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, which measures 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.

Column 1 of Table 3 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 show that father's occupation and caste networks both increase wages significantly and substantially in magnitude, but these effects are not significantly different in traditional occupations. Including all covariates together in Column 3 does not change the findings.

For women (Columns 4-6) we find much lower and imprecisely measured returns to education and experience in all occupations. For women, we again find large returns to fathers' occupation and caste networks which do not differ significantly between traditional and non-traditional

[11] The result also holds for women with a slightly smaller coefficient (see Column 3). Note that we condition on labor force participation, since home workers have no income data, which affects the sample size for women.

[12] 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.

occupations.[13]

Selection and productivity: Model implications

Our reduced form results show that traditional occupations are important for occupational choices. Yet the standard selection in a Roy model with uncorrelated occupational productivity cannot fully explain the wage patterns that we observe for workers in their traditional occupations.[14] In the uncorrelated Roy model, as emphasized by Young (2014), the first workers to enter an occupation have the highest occupation-specific productivity, so that the average occupation-specific productivity of a caste would decrease if preferences draw more of its workers into a given occupation. This finding is consistent with the empirical result that workers earn less in their traditional occupation than their caste mates in other (non-traditional) occupations (see Columns 1 and 3 of Table 2). Note that we find no significant difference in wages for workers in their fathers' occupations (Columns 1 and 3) since workers choose the occupation due to productivity gains from intergenerational learning (rather than preferences) which offset workers' negative selection on their occupation-specific productivity.

However, when comparing workers within an occupation we find that traditional workers earn more than outsiders in the same occupation (see Columns 2 and 4 of Table 2). This finding contrasts with the comparative static predicted by the uncorrelated Roy model. To understand this result, we show that traditional occupations have lower returns to general ability than "modern" (non-traditional) occupations. If workers differ in general ability, then high-ability workers sort a priori into modern occupations and low ability workers into traditional occupations. It follows that the marginal traditional workers who are drawn into their traditional occupation due to the utility boost have higher general ability than the average (non-traditional) worker in the same occupation. Due to the higher general ability, traditional workers can earn more in their traditional occupation than non-traditional workers in the same occupation, as we find in the data. Intuitively, our results indicate that some high ability individuals continue working in their low-return traditional occupations (e.g., agriculture, laundering, pottery) when, in the absence of the caste-occupation affinity, they might apply their skills more productively in high-return occupations (e.g., teaching, engineering, law). Guided by these empirical findings, we specify our general equilibrium occupational choice model to study the importance of caste identity on aggregate outcomes.

[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] 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.

Discrimination

We examine whether there is wage discrimination by gender or by castes' social ranking. To do so, we proxy castes' ranking by their categorization into other backward classes (OBC), scheduled castes (SC), and scheduled tribes (ST).[15] All specifications control for individual characteristics, including education and experience. Column 1 of Table 4 shows very large wage discrimination against women (-0.6 log-points) and much smaller effects for castes lower in the traditional hierarchy. With occupation fixed effects (Column 2), we find smaller estimates for women (albeit still large at -0.4 log-points) but larger and more precise estimates for lower-ranked castes (-0.1 log-points for OBCs, -0.2 for SCs and -0.2 for STs). The estimates are robust to controlling for father's occupation, caste networks, and a traditional occupation dummy (Columns 3 and 4).

5 Model

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

5.1 Model Setup

Individual Characteristics:

Individuals i differ in general ability, 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:

Individuals i belong to a caste k that affects their utility payoffs and choices via four channels. First, workers have a direct preference for working in their caste's traditional occupation, which we denote by τ_{ok}. Second, caste members can experience wage discrimination, which we denote by T_k. Third, workers receive productivity effects from their caste network, which we define as the share of all workers in an occupation who belong to the worker's caste. Last, caste affiliation can affect costs of schooling, which we denote by κ_k. These costs are measured in utils per year of schooling and can capture pecuniary costs (such as school fees or scholarships) 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

[15] 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.

[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.

notation, we denote the total productivity shifter from caste networks and father's occupation for an individual i in occupation o by ψ_{io}.

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 ability, which we denote by ρ_o and which capture the inherent skill-intensity of an occupation (engineering is for example more complex and skill-intensive than agricultural labor). We further allow occupations to vary in their amenities A_o, which capture that some occupations might be more pleasant than others and also that some occupations can have entry cost that are not directly measurable in the occupation's wages. In the Indian context, examples of such entry costs can include exams to enter government services 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 given by:

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

5.2 Education and Occupation Choices

The timing of the model is the following: Individuals live two periods, childhood and adulthood. At birth, they know their caste affiliation, their general ability α_i and their education cost shock η_i. Individuals first choose years of schooling s_i, which remain fixed during adulthood and is a component of workers' observable 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} is occupation-specific productivity, (α_i, β_i) measure 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 of choosing occupation o (net of the occupation-specific productivity shock). We provide the complete 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:

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

[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.

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

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

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 probability is equal to:

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

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

5.2.2 Education Choice

Caste affiliation matters for education choices by affecting the cost of education and 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. To capture both of these channels, we augment the standard Mincerian formulation by caste-specific education costs, so that we model children's education choices as:

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

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, capturing expected returns to schooling from higher wages during workers' years in the labor force. The second term of Equation 7 represents the cost of education, including both caste-specific shifters κ_k and idiosyncratic education cost shocks η_i. To facilitate estimation, 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. 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 ability, and the structure of occupational caste networks. These factors together determine the total amount of human capital that is supplied to each occupation. To derive human capital supply in each occupation, we first solve for workers' expected occupation-specific productivity π_{io} conditional on having chosen occupation o which is equal to:

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

where $\Gamma(\cdot)$ is the gamma function and where we used Assumption 1 to derive the closed form solution. This expression is standard in the uncorrelated Roy model and illustrates that selection on occupation-specific productivity π_{io} is negative within α-types: caste members have an affinity for their traditional occupation which increases their propensity of choosing the occupation ($P_{io|\alpha_i}$) and decreases their average occupation-specific productivity in the occupation. Next, we combine this measure with workers' other characteristics (cf. Equation 4) and we sum over all workers to express total expected human capital in occupation o as:

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

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 castes' occupational choice probability and average human capital is no longer guaranteed. If the utility of working in one's traditional occupation sufficiently attracts high α-types into low-return traditional occupations, then the average human capital of traditional workers in that occupation can be greater than that of outsiders due to their better α-skill composition.

5.2.4 Social Networks

We define caste-occupation networks as the share of all workers in an occupation that belong to a given caste. It follows that networks are endogenously determined by occupational choice probabilities in the following way:

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

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 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 experience a disutility from employing workers that belong to certain castes similar to Hsieh et al. (2019).

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 observed occupational choices and wages, and the second for the likelihood of observed educational choices. We first describe how we parametrize 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 allow the value of τ_{ok} to differ by 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

[18] Father's occupation might also have a direct impact on utility and could therefore be included in τ_{ok}. We tested this alternative specification and found this effect to be insignificant for males and slightly negative for women, perhaps due to 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 through functional form assumptions.

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.

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:

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

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:

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

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 transfer their occupation-specific knowledge to sons or daughters, and that social networks may be differentially 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 allow for wage discrimination based on caste hierarchy and gender in the following way:

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

We set $T_k = 0$ for homework because caste discrimination is unlikley within the home and because women may face systematic discrimination in all market occupations.

Education cost: We allow costs 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 where each element represents an occupational category.

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

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 value distributed with dispersion parameter σ_π; second, for general ability shocks α_i, 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 parameterizations at hand, we now turn to our maximum likelihood estimation. The estimation proceeds in two steps: the first for occupation choices and wages, and the second for education choices.

Occupation and wage likelihood

The occupation and wage likelihood function estimates the first set of parameters, which are $\Omega_{occ} = \left\{\tilde{\tau}, \tilde{\beta}, \tilde{\psi}, \tilde{\delta}, A_o, w_o, \rho_o, \sigma_\alpha, \sigma_\pi\right\}$. The likelihood that a worker i 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, so that:

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

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 (cf. 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):

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

The occupational choice and wage probabilities are both 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 ability 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,

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

an environment with high returns to skill in all occupations and a small variance of skills is observationally equivalent to one 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.

The wage component of the likelihood is not defined for home workers since they have no observable wage data. We therefore set it to 1 for home workers, using only their occupational choice data to estimate occupational parameters of homework. It follows that "wages" for home workers are not separately identified from amenities A_o so that we normalize $w_{home} = 1$. Crucially, we do not endogenize the homework "wage" in counterfactuals and we measure output only from market workers.

Education likelihood

With the first set of parameters at hand, we 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 $\hat{y}_{io}$ and occupations $\hat{o}$ as well as on parameters Ω. To characterize these education cost shocks we rearrange the first order conditions of the education choice problem (cf. 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. When solving for the likelihood, we impose the constraint that the second-order conditions of the education choice problem must be negative at the optimal education level to ensure that we derive education choices that 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

[21] Education choices are 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.

the education cost κ in the education likelihood, but non-linear in other parameters. We therefore implement the estimation in two steps and bootstrap the standard errors to account for this 2-stage process, clustering at the PSU level.

6.3 Backing out production parameters

Last, we need to determine the parameters from the CES production technology, which are occupational intensity Z_o, total factor productivity A, and the elasticity of substitution σ across occupations. We calibrate σ to the literature (setting it to 2/3) and we compute the other parameters by matching the model's optimality conditions to the data. Dividing firms' first order conditions across two occupations yields the following expression:

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

where we can compute relative occupational shares Z_o by using our estimated wage rates w_o and by constructing human capital in each occupation Θ_o from the data. The level of occupation shares Z_o is identified because they have to sum to one across all occupations. Using the estimates of σ and Z_o, we then rearrange firms' first order conditions to infer total factor productivity A from:

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

7 Results

We now discuss the results from our maximum likelihood estimation and from our counterfactuals.

7.1 Structural Parameters

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

The first two columns of Table 5 present our estimates of workers' preferences for working in their traditional occupation–the key focus of our paper. To provide an interpretation of the estimated parameter values, let us consider their effects on occupational choice probabilities, which are shifted by $\exp(\sigma_\pi \tau_{io})$, as shown in Equation 5. The coefficient on traditional occupation (0.18) 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 for whom the combined coefficient from row 1 and 2 implies a 12.4 percent greater attraction to their traditional occupations. 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

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

is significantly stronger for OBC castes, not significantly different for scheduled castes, and significantly weaker for 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 of education and experience into general human capital units β_i. These coefficients are analoguous to the standard Mincer coefficients. Note that the values capture the baseline returns for any occupation, while actual returns are adjusted by the occupation-specific return to general human capital ρ_o (shown in Table A3). 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). Experience has positive but diminishing returns for men, while it 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 percent increase in the share of workers of an occupation who are of the same caste as the respondent is associated with a 8 percent 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 estimates of 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 OBCs and SCs–indeed wages appear marginally higher for both groups and STs 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 these policies and negative discrimination.

Column group 3 presents the structural coefficients that determine the cost of education. We find that education costs are convex and negative for low years of schooling 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 SCs and STs. These non-pecuniary rewards (i.e., negative costs) of receiving low education can reflect social stigma, returns on the marriage markets, or compulsory elementary schooling. Schooling choices further depend on idiosyncratic education cost η_i and 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. We display the full parameter vectors in Appendix Table A3. 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

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

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 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.

Distributional parameters: Table 6 shows the estimates of distributional parameters. We find that the dispersion of occupational-specific productivity σ_π, of general ability σ_α, and of education cost σ_κ is higher for women than for men. These gender differences could be driven by forces outside our model such as household formation or fertility.

7.2 Counterfactual Results

Our counterfactual analysis explores how the Indian economy would differ if caste identity was not linked to traditional occupations. We evaluate the effects of all three channels that link castes to their traditional occupations. First, we remove in all counterfactuals the direct attachment to traditional occupation τ_{io}, 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 as their father, which can keep workers in their traditional occupations. We analyze the importance of this channel in a counterfactual in which we remove any correlation between traditional occupation and the distribution of father's occupation. To do so, we use a dataset at the occupation$\times$individual level and we regress an indicator for father's occupation on an indicator for traditional occupation and a constant. 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 caste-occupation networks.[24]

We present our results in Table 7. First, we fix caste-occupational networks (Panel A) and the distribution of fathers' occupations (Columns 1 and 2). These counterfactuals answer the question, "what would happen if the current generation of workers stopped valuing their traditional

[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.

occupation?" Column 1 of Panel A further holds workers' education constant, so that we evaluate the direct effect of eliminating traditional occupation preference τ_{io} when only occupational choices and wages adjust endogenously. The impact on the aggregate economy is extremely minor with an increase in aggregate output by 0.06 percent and in output per worker by 0.3 percent. In addition, labor force participation decreases as some workers leave their traditional occupations to enter homework. In Column 2, we 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 preference parameters $\tilde{\tau}$ is small relative to the variation in other structural parameters, in particular relative to amenities A_o and relative to productivity effects from networks and fathers' occupation ψ_{io}. The basic structure of the economy therefore remains relatively unchanged as shown in Panel A.ii of Table 7: employment shares drop by at most 0.72 percentage points for any occupation, even if the share of traditional workers drops by up to 5.35 percentage points for the most affected occupation and by 11.6 percent (1.2 percentage points) in the aggregate. This finding implies that traditional workers simply get replaced by other (similar) workers, thus keeping the occupational structure and aggregate output similar.[25] Despite the trivial aggregate effects, we see improvements in workers' selection based on their occupation-specific productivity π_{io} and their general ability (α_i, β_i), as individuals with higher general ability increasingly select into occupations with high returns to ability ρ_o. 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 intergenerational learning and caste networks.

In Column 3 of Panel A, we eliminate one more channel of historical persistence by removing the correlation between fathers' occupations and traditional occupations. Aggregate output now decreases by 2.98 percent. A part of this drop is driven by a 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. Changes in output per worker are therefore much smaller but still slightly negative with a decrease by 0.005 percent. Effects are negative because workers now have to choose between either their father's occupation or their traditional occupation where caste networks (which we hold constant) are strongest. Overall, the productivity losses from less intergenerational learning and less sorting towards strong caste networks are larger than the gains from workers' improved selection based on their individual characteristics. We now find larger distributional effects for castes and occupations: the most affected caste sees a 54 percent decrease in average income, the most affected occupation loses 21 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).

We then implement the same counterfactuals while allowing caste networks to adjust endoge-

[25] Banerjee et al. (2013) find similar results for a different aspect of caste identity: individuals' preference to marry within caste. The authors show that removing the preference for same-caste spouses significantly reduces the share of intra-caste marriages, but has only minor effects on marriages and houshold compositions along other characteristics.

nously. 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 individual choices and market clearing. We then compare all counterfactuals to this baseline steady state. There is a possibility of 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 numerical approach that aims at identifying 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).

Panel B of Table 7 presents the effects of each counterfactual with endogenous networks. In Column 1 we hold again fathers' occupations and education constant and we find a decrease in aggregate output by 0.15 percent, a reduction in labor force participation by 0.3 percent and an increase in output per worker by 0.17 percent. Removing castes' occupational affinity now leads to weaker caste-occupation networks, which reduces output and productivity compared to the constant network case (Column 1, Panel A). In Column 2 we further allow education to adjust endogenously, which increases output by 0.76 percent, output per worker by 1.1 percent and schooling by 0.7 percent. Compared to Panel A, gains with endogenous networks are now larger, because workers reallocate their human capital more efficiently and choose more education. We again find that the basic structure of the economy changes little, even if workers are less likely to choose their traditional occupation (see Panel B.ii).

Last, in Column 3 of Panel B we eliminate the influence of traditional occupation via parental occupation. Effects are now remarkably large with an 8.1 percent decrease in output, a 3.2 percent decline in labor force participation and a 5.1 percent drop in output per worker. These results highlight that castes' ties to their traditional occupation–either direct ties via preferences or indirect ones via parental occupations–play an important role in organizing castes into strong occupation networks. This coordinating element is essential because individuals do not internalize the large externality of their occupational choices on caste networks. Removing these coordinating elements and allowing networks to adjust endogenously therefore leads to occupational networks that are only weakly clustered at the caste level, which lowers output and productivity. Consistent with this, we find a large decline in the aggregate share of traditional workers which decreases by 24 percent (4 percentage points). For the most affected occupation, the share of traditional workers decreases by 10.6 percentage points and for the most affected caste by 27.5 percentage points. Despite the negative aggregate effects, we find improvements in education (2 percent) and in the selection of workers based on their individual characteristics–without strong caste networks in low-return occupations, agents invest more in education and are more likely to pursue occupations aligned with their comparative advantage. This leads to larger changes in the occupational distribution with occupations contracting by up to 33 percent and others expanding by up to 9 percent in

[26] To do this, we start from the baseline human capital distribution and caste networks and we perform only minor updates, changing exogenous parameters in several small steps, when solving for the new equilibrium.

human capital.

To document effects of inequality more systemically, we use Growth Incidence Curves, proposed by Ravaillon and Chen (2003) and popularized by Milanovic (2016), to compare the dispersion of human capital in the baseline and counterfactual economies. We focus on human capital, instead of income, because it is defined regardless of labor force participation. We define individuals' human capital broadly as occupation-specific productivity, general ability, education and experience as well as productivity effects from social networks and parental learning. Figure 2 presents the Growth Incidence Curves, where the x-axis ranks workers by percentile of their baseline human capital and the y-axis shows the mean growth rate in human capital for each percentile.[27]

Figure 2 shows that workers in the middle of the baseline human capital distribution gain most, while workers with initially low human capital lose. Figure 2a presents the effects with constant social networks which display a particularly clear U-shaped pattern. Workers with the lowest initial human capital, 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 weaker networks and the reduction of paternal learning in traditional (low-skill) occupation. The highest human capital workers are already in high-skill occupations and have little scope to increase their education (since education costs are convex). In addition, these workers might experience more competition from new entrants into high-skill occupations, which can decrease wages and make additional education investments less attractive. Figure 2b shows the effects with endogenous social networks which display a similar pattern. Workers with lowest initial human capital lose most due to weaker caste networks in their traditional occupations. With endogenous networks, workers with high initial human capital now benefit more, as their social networks become stronger in high-skill occupations. These results confirm the findings of Munshi and Rosenzweig (2006) who show that caste networks are important and particularly benefitial for low-skill workers in their 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. Occupational identity has a major effect upon career choice–in India certain occupations are still composed primarily of individuals "born" into that occupation, and the average person is more than three times as likely to enter their traditional occupation than any other. It follows that caste-based occupational identities continue to affect the selection of occupations well into the modern era.

However, we find that these large effects of caste identity on occupational choices have only a small impact on the overall efficiency of the economy. This is because three forces diminish the distortions of caste-based choices on the macro-economy. First, working in a parent's occupation

[27]Note that computing the mean growth rate is different from computing the growth rate of the mean, and has preferable properties as discussed in Ravaillon and Chen (2003).

increases productivity and castes' traditional occupations are highly correlated with their parental occupations. Hence, we find that many workers would continue to work in their traditional occupations, even if the non-pecuniary utility of doing so is removed, because they can learn occupation skills from their fathers. Second, the clustering of castes into occupations generates positive social network effects that partially compensate for the misallocation of human capital. Even if individuals may be working in the "wrong" occupations, by doing so they increase the productivity of all their caste-mates in that occupation. Third, the misallocation of human capital caused by traditional occupations is primarily limited to the reallocation of low-skill individuals between different 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".

We find larger effects when we allow social networks to 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-occupation level, lowering network-based productivity spillovers. Overall, these factors lead to a reduction in market output, despite improvements in education and human capital allocation. These results highlight the need of taking a broad view when studying the economic importance of frictions. While the direct effects of occupational preferences are small, their indirect effects via social networks and intergenerational learning are much more important.

An important limitation of this study, inherent in the revealed preference approach to occupational identity, is that we can only identify relative and not absolute values of occupation-linked utility. We can therefore not distinguish between a scenario in which individuals receive positive utility from working in their traditional occupation, versus an alternative in which they receive negative utility from all other occupations. With this caveat in mind, we limit our counterfactual analysis to study effects on income and inequality, leaving the evaluation of welfare effects to future work that uses a different methodology.

Our analysis suggests a possible explanation for the remarkable persistence of occupational identities in the 21st century. If the static economic costs are mild, but individuals receive substantial utility from conforming with social norms, then these norms can persist over long periods. This may be one reason why castes' 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, that the main costs of identity frictions may be dynamic and occur over the course of structural transformation. For individuals who are 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*.

Banerjee, A., E. Duflo, M. Ghatak, and J. Lafortune (2013). Marry for what? caste and mate selection in modern India. *American Economic Journal: Microeconomics 5*(2), 33–72.

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 (2019, September). 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

Notes: 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 respondents and occupations. The outcome variable is equal to 1 for respondents' chosen occupation and 0 for all other occupations. The variable "Occ. is caste's trad. occ." indicates that an occupations is traditionally performed by the respondent's caste (if any), as defined in Section 3. Caste-occupation networks are equal to the jackknifed ratio between the number of respondents' caste-mates in an occupation divided by the number of all workers in the occupation. The scheduled caste (SC) dummy indicates whether the respondent's reported caste belongs to 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 trad. occ.	-0.238***	0.128***	-0.191**	0.115**
	(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 trad. occ.×SC	0.172***	-0.065	0.317**	-0.083***
	(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

Notes: This table reports results of regressing 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, trimming the 1st and 99th percentiles. The variable "Occ. is caste's trad. occ." indicates that an occupations is traditionally performed by the respondent's caste (if any), as defined in Section 3. Caste-occupation networks are equal to the jackknifed ratio between the number of respondents' caste-mates in an occupation divided by the number of all workers in the occupation. The scheduled caste (SC) dummy indicates whether the respondent's reported caste belongs to 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

Notes: This table reports results of regressing 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 variable "Occ. is any trad. occ" indicates whether an occupation is traditional for *any* caste, as defined in Section 3. Caste-occupation networks are equal to the jackknifed ratio between the number of respondents' caste-mates in an occupation divided by the number of all 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

Notes: This table reports results of regressing 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 variable "Occ. is caste's trad. occ." indicates that an occupations is traditionally performed by the respondent's caste (if any), as defined in Section 3. Caste-occupation networks are equal to the jackknifed ratio between the number of respondents' caste-mates in an occupation divided by the number of all 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}) (1)		General human capital ($\tilde{\beta}_i$) (2)		Costs of education (κ_i) (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 ($\tilde{\psi}_{io}$) (4)		Labor force discrimination ($1 - T_{io}$) (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)

Notes: Parameters are estimated by maximum likelihood as 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)

Notes: 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

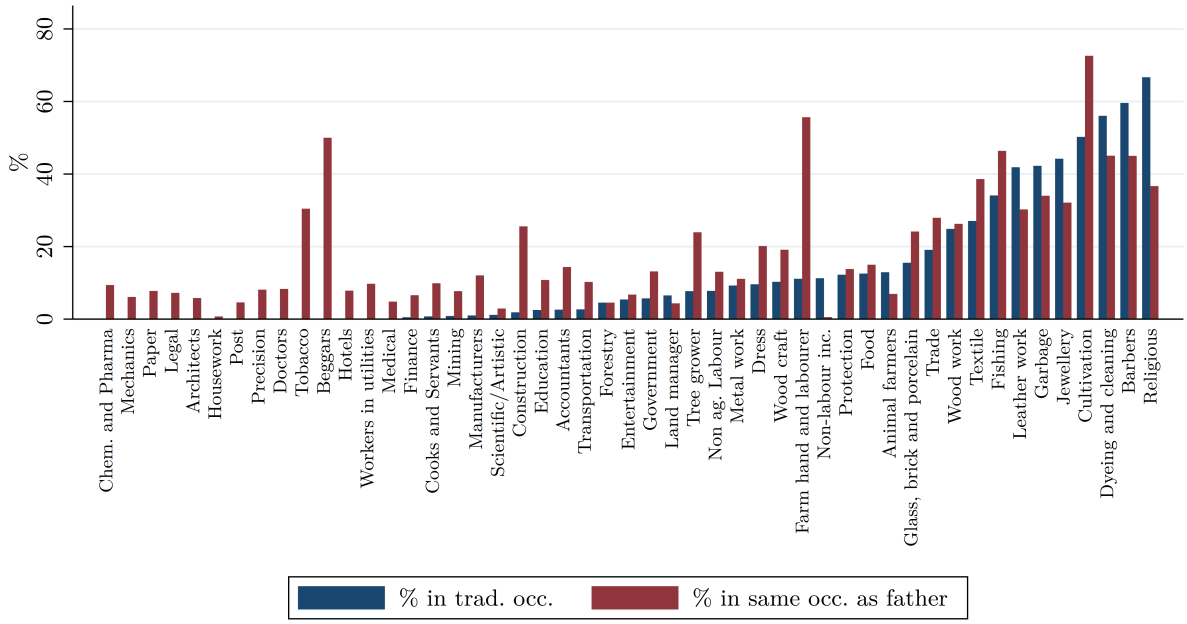
	No occupational identity + endogenous wages (1)			(1) + endogenous education (2)			(2) + parental occupation orthogonal to trad. occ. (3)		
Panel A: Estimated Social networks									
i. Aggregate outcomes (percent changes from baseline)									
Market Output	0.062			0.346			-2.980		
Output per worker	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	-0.09	-0.03	0.55	-0.28	-0.05	0.55	-1.60	-1.03	6.89
Occupation: human capital	-1.56	0.17	0.34	-1.29	0.5	1.18	-20.56	0.1	2.43
Occupation: employment share (pp)	-0.68	0.003	0.19	-0.72	0.004	0.2	-1.99	0.003	2.07
Occupation: trad. worker share (pp)	-5.2	-0.42	0.00	-5.35	-0.42	0.00	-12.41	-1.02	0.00
Caste: % workers in trad. occ. (pp)	-5.68	-0.24	0.34	-6.08	-0.25	0.34	-23.71	-0.79	0.76
Caste: total income	-0.88	-0.004	1.06	-0.79	0.07	3.38	-53.52	-0.42	13.54
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	-0.26	-0.18	0.55	-0.97	0.004	0.73	-5.47	-3.09	11.08
Occupation: human capital	-1.79	0.39	0.64	-1.42	0.77	3.76	-32.97	1.03	8.75
Occupation: employment share (pp)	-0.63	0.004	0.24	-0.66	0.01	0.22	-2.04	0.01	2.42
Occupation: trad. worker share (pp)	-3.64	-0.3	0.00	-3.86	-0.3	0.00	-10.65	-0.7	0.00
Caste: % workers in trad. occ. (pp)	-5.93	-0.24	0.41	-6.17	-0.24	0.44	-27.48	-0.77	1.65
Caste: total income	-0.97	-0.11	1.0	-1.15	0.13	3.48	-56.12	-2.26	5.50

Notes: Results in Panel A are relative to a baseline economy that is simulated from the estimated parameter values and the caste network data used in the estimation. Results in Panel B are relative to a baseline economy with endogenous caste networks for which we solve conditional on all estimated parameters. All counterfactuals use posterior 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.

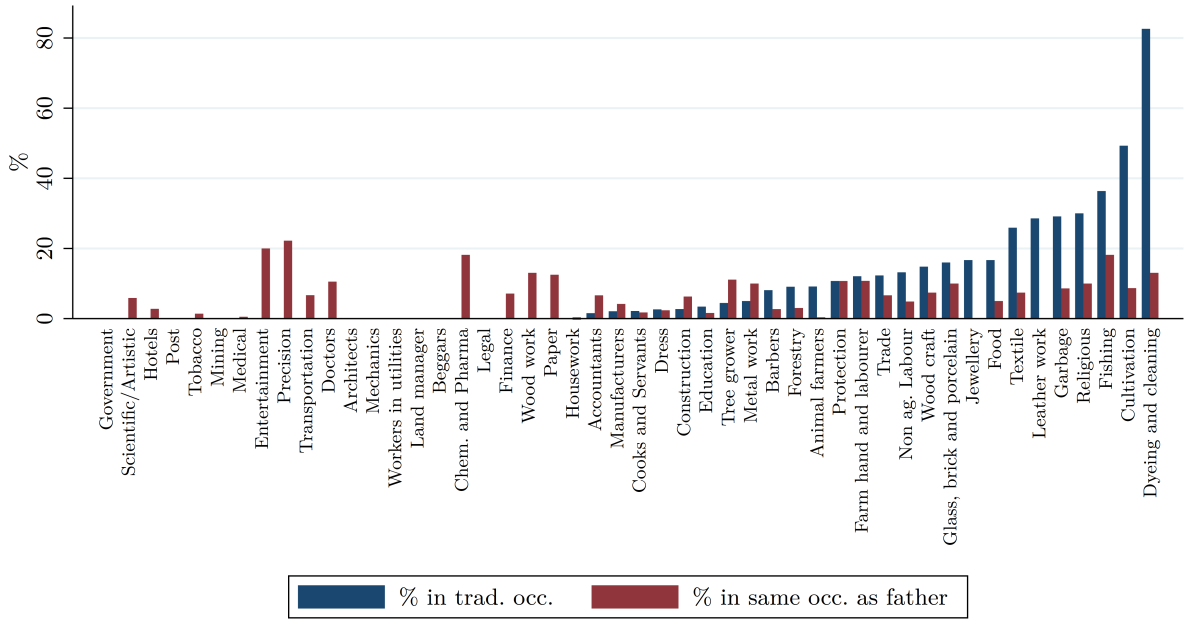
10 Figures

Figure 1: Occupational Composition: Traditional and Parental Transmission

(a) Male Workers



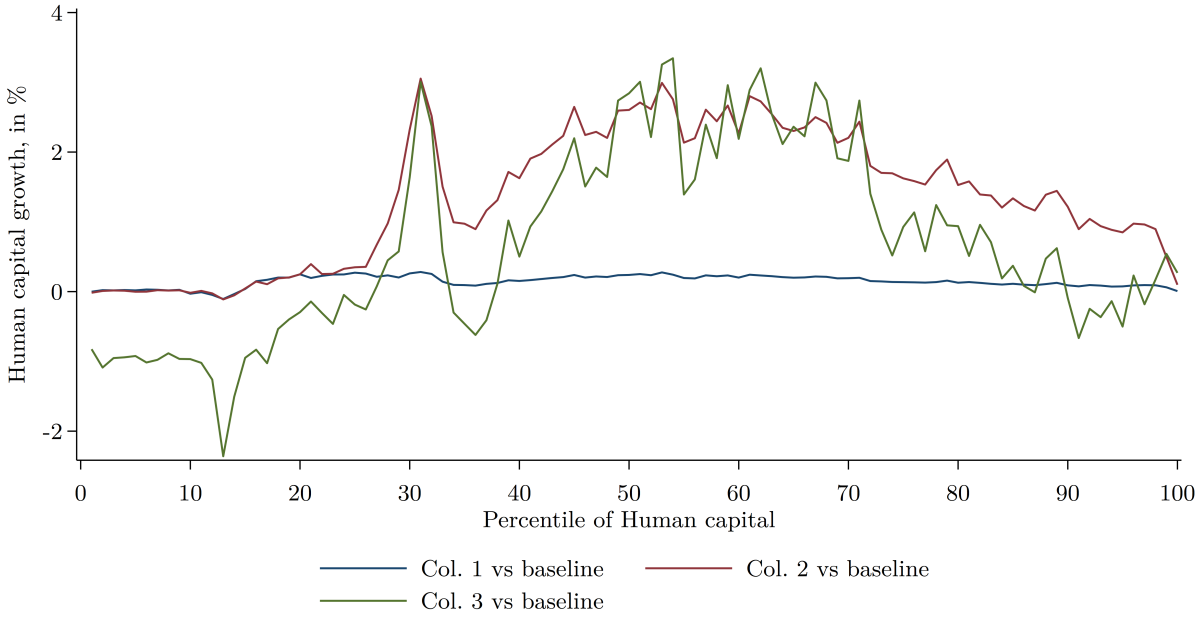
(b) Female Workers



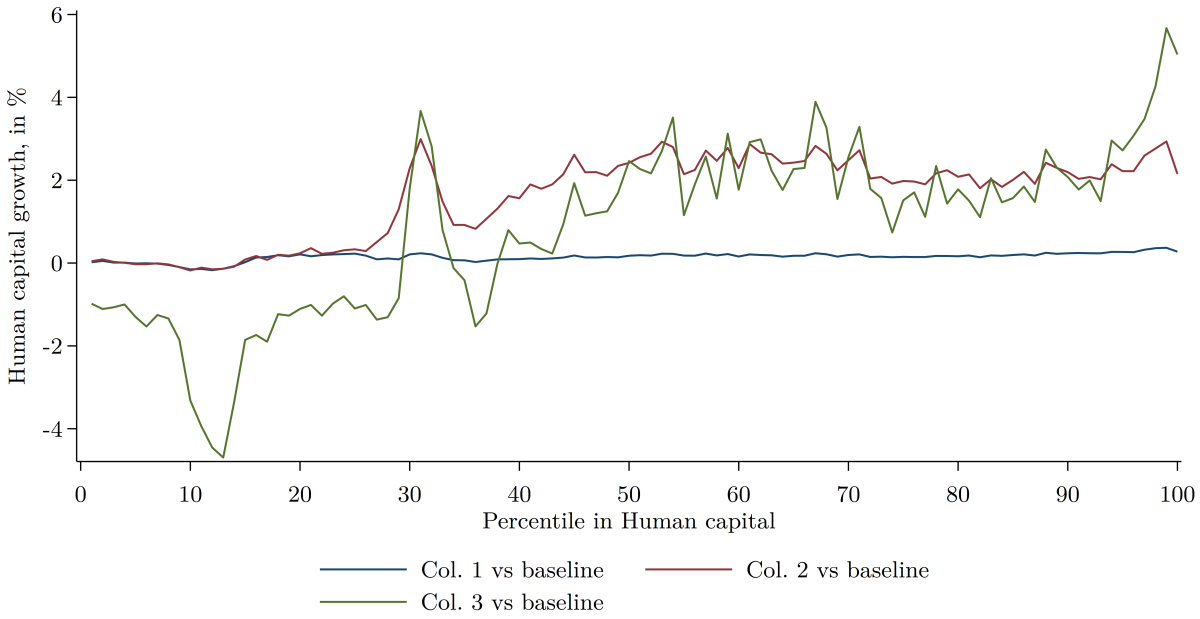
Notes: Panel (a) documents the share of male workers in each occupation who work in their jati's traditional occupation (blue bars) and the share who works in their father's occupation (red bars). Panel (b) does the same for women.

Figure 2: Counterfactuals: Human Capital Growth Incidence Curves

(a) Counterfactuals with Exogenous Social Networks (see Table 7, Panel A)



(b) Counterfactuals with Endogenous Social Networks (see Table 7, Panel B)



Notes: This figure shows Growth Incidence Curves in human capital to document the distributional effects of our counterfactuals. The x-axis ranks workers by percentile of their baseline human capital. For each counterfactual, the y-axis shows the mean growth rates of human capital for workers in each percentile. Panel (a) shows the effects of counterfactuals with exogenous caste-occupation networks and Panel (b) with endogenous networks.

Appendix

A1 Data Appendix

The IHDS records household and individual income from a variety of different income sources. We restrict our sample to individuals aged 18 to 60 and drop full-time students and unemployed individuals. The survey provides time spent and income earned from most occupations at the individual level, however, we have to make assumptions to derive individual income for some occupations.

First, income is only reported at the household level for some occupations such as household businesses. For these occupations, we know how much time each household member works in them, but we do not know individuals' productivity, so that we attribute the same hourly wage to each individual.

Second, the IHDS does not record the time spent in animal care but it records how many animals of each type are owned by the household, and which household members take care of the animals. To derive information on hours worked, we use an additional household survey, the REDS 2006, which is representative for rural India. The survey contains information on time spent on animal care by household members and on the number and types of animals owned by a household. We predict the time that IHDS household members spent 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 squared term.

Third, the IHDS does not provide time spent in some occupations such as money lending and the rental of land. We did not find alternative data sources that would allow us to infer this information, so that we instead use information on time spent and income earned by the same individual in other occupations. For individuals that record income and hours worked for another occupation, we compute the time spent in money lending or land rentals 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 almost 34 percent of the respondents in our sample report their "primary activity" as housework, who are almost entirely women. However, many of these women indicate that they also spend many hours working for income. We therefore designate a respondent's main occupation only as housework if she works less than the median number of hours in all other reported occupations (or if she reports no other work). Otherwise we assign her to the occupation in which she earns the highest income. With this procedure, the final sample classifies 12.3 percent of respondents as home workers.

Fifth, because many individuals report multiple sources of income, we classify individuals' occupation as the activity in which they earn highest income and spend most time. If these two definitions are different, then we choose the occupation that the survey identifies as the "main activity" of the household, and then the activity in which individuals' spend most time.

Sixth, we trim hourly wages at the 1st and 99th percentiles. Results are robust, and in fact stronger, when we instead trim at the 0.1st and 99.9th percentiles.

A2 Additional empirical results

A2.1 Robustness of occupational choice

We perform a variety of robustness checks to confirm the importance of traditional occupations for occupational choices. In Table A1, we show that our main occupational choice analysis (presented in Table 1) is 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., an occupation traditionally attached to high castes) and the individual belongs to a scheduled caste, and whether the occupation is agricultural and the individual inherited land. Finally, we allow for intrafamilial transfers of human capital, in addition to direct transfers from father to child, by controlling for whether individuals work in the same occupation as their uncle. None of these additional controls change our finding that individuals are more likely to work in their jati's traditional occupation.

A2.2 Robustness to imputation of parental occupation

The IHDS data contains information on the occupation of the father of the household head but no information on the mother's occupation. As a consequence, we have very little information on father's occupation for women. In our main analysis, we replace missing data on father's occupation with probabilities based on the data of fathers' occupations for individuals in the same caste. Here we test how this imputation may affect our results using the National Election Survey (NES), which contains information on parents' occupations for all respondents. This data was collected by the Center for the Study of Developing Societies (CSDS) in 2009 and 2014 based on a sample of registered voters from electoral rolls and contains detailed information about the occupations of respondents and both of their parents. The dataset also reports respondents' jati names, which have been cleaned by the CSDS but which are not reported verbatim as in the IHDS and DHS surveys (our main datasets). In addition, the jati categorization, while detailed, is still much more aggregated than the one that we use in the main part of our paper. Finally, the dataset does not provide precise income data, so we cannot use it as our main data source.

We combine the 2009 and 2014 NES samples and keep only the respondents aged 18-60 for which the jati is identified. We use this dataset to reproduce the occupational choice regressions presented in Table 1 which relied on IHDS data. In the NES data, we now directly observe father's occupation for women instead of imputing it and we observe mother's occupation for women and men. To compute jati-occupation networks, we combine the NES data with both IHDS and DHS. To make jati categorization comparable across the three surveys, we aggregate several jatis to the broader categories provided in the NES. Table A2 presents the results of our regressions. The coefficients on "Occ. is father's occ." are similar to our main results for both genders, indicating

that our imputation method is of little consequence. Including mother's occupation does not affect the coefficient on father's occupation for men, but reduces it for women. The coefficient on jati-occupation network is substantially reduced for both genders, which we attribute to the fact that our measure of jati in these regressions is much more aggregated than the one we use in our main regression results.

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., $s + T$), so that:

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

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:

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

where $\bar{r} = \frac{e^{-rs}}{r} \left(1 - e^{-rT} \right)$ is the discount factor that incorporates years spent in school, and where we define $\bar{\beta}_i = \exp \left(\tilde{\beta}_s s_i + \bar{\beta}_x \right)$ where $\bar{\beta}_x$ is the pre-employment expected value of experience which is 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 productivity shocks π_{io} are not known at this time, so 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 (left hand side of the equation) with marginal returns (right hand side of the equation). Education costs depend on the direct costs (κ and η) and the opportunity cost from foregone income. Total returns to education depend on the generic return 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 $\{\tilde{\beta}, \tilde{\psi}, \rho_o\}$ determine worker i's productivity in each occupation.[28] We assume that entrepreneurs experience a disutility from hiring certain castes in certain occupations, which we denote by δ_k. This disutility generates wage discrimination T_k which affects castes' effective occupational wage rate per human capital unit. Exogenous parameters that characterize the production function are total factor productivity A, occupation shares Z_o and the elasticity of substitution between occupations σ. Individuals' utility depends on their education cost κ_k, occupation amenities A_o, and their preferences for working in their traditional occupation τ_{ok}. Last, the parameters σ_π, σ_α and σ_η respectively characterize the dispersion of the idiosyncratic productivity shocks π_{io} , general ability shocks α_i and education cost shocks η_i. We denote the full set of exogenous parameters by: $\Omega = \left\{\tilde{\beta}, \tilde{\psi}, \rho_o, \delta_{ok}, A, Z_o, \sigma, A_o, \tau_{ok}, \kappa_k, \sigma_\pi, \sigma_\alpha, \sigma_\eta\right\}$. Given these exogenous parameters Ω, the equilibrium of the economy is characterized by:

1. Occupational choice probabilities P_{io} that are consistent with individuals' utility maximization (c.f. Equation 5).

2. Education choices s_i that are consistent with individuals' utility maximization (c.f. Equation 7).

[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.

3. Human capital demand Θ_o in each occupation that is consistent with firms' profit maximization (c.f. Equation 10).

4. Wage discrimination that exactly offsets entrepreneurs' disutility of hiring certain castes, so that $T_k = \delta_k$ which makes entrepreneurs indifferent between hiring workers from any caste.

5. Wage rates per human capital unit w_o tha clear labor markets in each occupation, ensuring that human capital demand equals human capital supply in each occupation (c.f. Equations 10 and 9).

6. Good market clears, so that total consumption equals total output.

A3.4 Derivation of the wage distribution and likelihood

Here, we derive the likelihood functions for observed occupations, wages, and schooling levels. We proceed in two steps: first, we derive the distribution of occupation-specific productivity shocks conditional on having chosen an occupation. Second, we build on this result to derive the distribution of workers' income conditional on their occupational choice.

Let V_i^* be the maximum utility of a worker who chooses occupation o^* before his occupational choice:

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

Assumption 1 states that $\log\left(\pi_{io}\right)$ is Gumbel distributed, which implies that the maximum utility level V_i^* is also Gumbel distributed, so that:

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

which corresponds to the CDF of the Gumbel distribution 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 productivity shocks π_{io}^* for individuals who have chosen occupation o:

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

From this expression, we see that productivity shocks in the chosen occupation π_{io}^* are Frechet

distributed with the mean being equal to the inverse of the occupational choice probability P^*_{io}. We now use this result to derive the distribution of workers' income y^*_{io} in their chosen occupation. Recall that earnings in our model are given by: $y_{io} = (1 - T_k)\, w_o\, (\alpha_i \beta_i)^{\rho_o}\, \psi_{io} \pi_{io}$, so that:

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

Last, we take the derivative to obtain the PDF of workers' income in their chosen occupation:

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

A4 Additional counterfactual results

A4.1 Algorithm

We solve for our counterfactuals with a fixed point algorithm. We first modify parameters (or model objects) 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 until they are consistent with our equilibrium definition.

A4.2 Counterfactual results with weaker network effects

As a robustness check, we implement all counterfactuals with weaker network effects to assess the sensitivity of our results to these parameter estimates. This exercise addresses the potential concern that our estimates of the productivity effects from social network effects may be biased. It is for example possible that unobservable characteristics make certain castes more productive in certain occupations–attracting more caste members to these occupations and increasing their wages–which we could not distinguish from network effects. We therefore recompute all counterfactuals with the strength of network effects reduced by half, formally dividing $\tilde{\psi}_2$ and $\tilde{\psi}_4$ by 2. Table A4 shows that our results with weaker network effects are qualitatively similar and also roughly similar in magnitude. An exception is the counterfactual that allows for endogenous networks and removes the correlation between one's father's and one's traditional occupation (Panel B, Column 3). With weaker network effects, this counterfactual leads to substantially smaller losses in output and output

per worker. However, our main qualitative results–that effects of removing links between castes and their traditional occupations are small and in some cases even negative–remain unchanged.

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.×				0.007***		0.007***
land inherited				(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

Notes: 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 respondents and occupations. The outcome variable is equal to 1 for respondents' chosen occupation and 0 for all other occupations. The variable "Occ. is caste's trad. occ." indicates that an occupations is traditionally performed by the respondent's caste (if any), as defined in Section 3. Caste-occupation networks are equal to the jackknifed ratio between the number of respondents' caste-mates in an occupation divided by the number of all workers in the occupation. The scheduled caste (SC) dummy indicates whether the respondent's reported caste belongs to 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: Traditional Occupation and Occupational Choice: NES data

	Probability of occupational choice				
	(1)	(2)	(3)	(4)	(5)
	A. Male (N =446,918)				
Occ. is caste's trad. occ.	0.055***	0.020***	0.020***	0.021***	0.021***
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
Occ. is father's occ.		0.541***	0.541***	0.541***	0.514***
		(0.009)	(0.009)	(0.009)	(0.009)
Caste-occ. network			-0.4777***	-0.480***	-0.489***
			(0.093)	(0.093)	(0.093)
Occ. is caste's trad. occ. * SC				-0.005	-0.005
				(0.006)	(0.006)
Occ. is mother's occ.					0.120***
					(0.007)
	B. Female (N =411,160)				
Occ. is caste's trad. occ.	0.023***	0.008***	0.008***	0.007***	0.005***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)
Occ. is father's occ.		0.250***	0.250***	0.250***	0.123***
		(0.011)	(0.011)	(0.011)	(0.009)
Caste-occ. network			-0.043	-0.041	-0.090*
			(0.050)	(0.051)	(0.047)
Occ. is caste's trad. occ. * SC				0.004	-0.001
				(0.007)	(0.005)
Occ. is mother's occ.					0.529***
					(0.014)
Individual FE	Yes	Yes	Yes	Yes	Yes
Occ. FE	Yes	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 2009 and 2014 NES. The dataset contains all unique combinations of respondents and occupations. The outcome variable is equal to 1 for respondents' chosen occupation and 0 for all other occupations. The variable "Occ. is caste's trad. occ." indicates that an occupations is traditionally performed by the respondent's caste (if any), as defined in Section 3. Caste-occupation networks are equal to the jackknifed ratio between the number of respondents' caste-mates in an occupation divided by the number of all workers in the occupation. The scheduled caste (SC) dummy indicates whether the respondent's reported caste belongs to the state-level list of scheduled castes.
Standard errors clustered at the PSU (constituency) level. $*p < 0.10, **p < 0.05, ***p < 0.01$

Table A3: Occupation-level Structural Parameters

	Occupation skill-wage ($\ln w_o$) (1)	Occupation Amenity ($\ln A_o$) (2)	Returns to skill ρ_o (3)
Non-labor 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)
Agricultural Laborers	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)
Plantation; 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)
Laborers, 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 A3: 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)

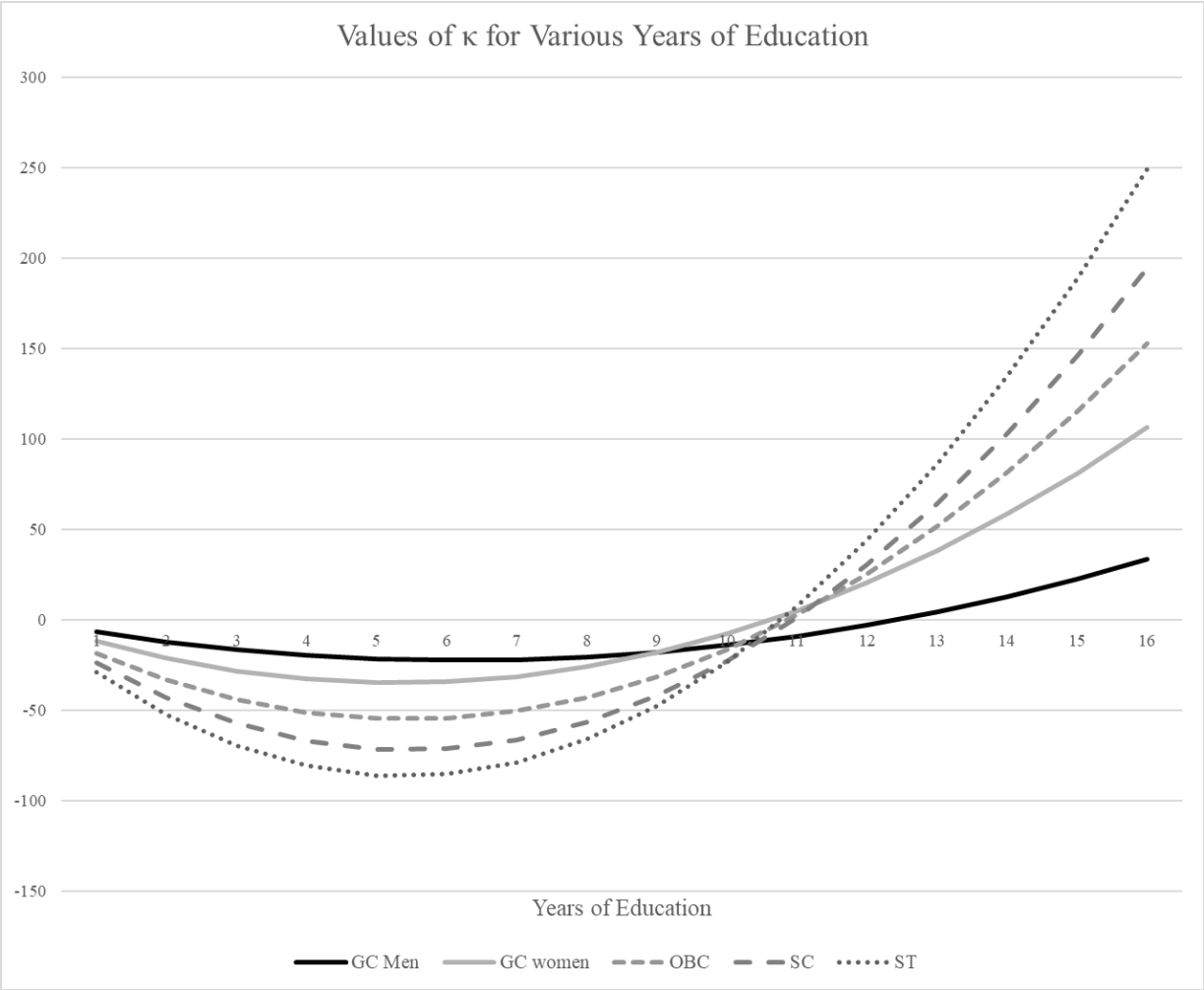
Table A4: **Effects of Removing Occupational Identity: Weaker network effects**

	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	-0.09	-0.023	0.3	-0.26	-0.04	0.42	-1.51	-0.95	6.30
Occupation: human capital	-0.82	0.16	0.34	-0.92	0.47	1.13	-19.06	0.09	2.30
Occupation: employment share (pp)	-0.67	0.003	0.18	-0.71	0.004	0.18	-1.92	0.002	1.96
Occupation: trad. worker share (pp)	-4.29	-0.39	0.00	-4.39	-0.4	0.00	-9.46	-0.93	0.00
Caste: % workers in trad. occ. (pp)	-5.68	-0.24	0.34	-6.08	-0.24	0.34	-23.69	-0.74	0.76
Caste: total income	-0.88	-0.003	1.06	-0.79	0.07	3.38	-53.49	-0.39	13.60
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	-0.06	0.96	-0.27	-0.07	0.81	-1.49	-1.03	6.20
Occupation: human capital	-2.82	0.2	0.38	-2.15	0.47	1.06	-18.90	0.34	2.18
Occupation: employment share (pp)	-0.67	0.003	0.17	-0.71	0.004	0.18	-1.87	0.002	1.9
Occupation: trad. worker share (pp)	-4.63	-0.4	0.00	-4.70	-0.39	0.00	-9.93	-0.88	0.00
Caste: % workers in trad. occ. (pp)	-5.69	-0.23	0.33	-6.09	-0.23	0.34	-23.66	-0.68	0.75
Caste: total income	-0.90	-0.01	1.05	-1.31	0.05	3.41	-53.32	-0.4	13.57

Notes: 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 education cost κ



Notes: This figure presents the cost of education κ by years of schooling for general caste men and women, OBCs, SCs, and STs.