# Is Better Access to Markets a Substitute for Educated Parents? Evidence from Rural India[*]

Reshad N. Ahsan[†]
University of Melbourne

Arpita Chatterjee[‡]
University of New South Wales

Shahe Emran[§]
IPD, Columbia University

September, 2018

*Preliminary version. Please do not cite or upload online.*

**Abstract**

This paper focuses on a particularly important channel, human capital investment, through which domestic market integration may affect intergenerational economic mobility. Using household data from India Human Development Survey, our empirical analysis focuses on the rural households. We use two sources of plausibly exogenous variations: (1) historical road infrastructure, and (2) crow-fly distance to the Golden Quadrilateral road network. The estimates from the OLS and 2SLS regressions for the average effect show that better market access improves childrens schooling attainment. More importantly, the interaction of fathers education with market access is negative; better market access thus works as a substitute for a higher educated father. The estimates from IV quantile regressions using the control function approach show that better market access weakens intergenerational persistence the most for the children in the middle of the conditional distribution.

*Keywords*: Market Access, Intergenerational Mobility.
*JEL Codes*: F14, F16, J62

[*]We thank Shoumitro Chatterjee for helpful comments and discussions.

[†]Department of Economics, Level 3, FBE Building, 111 Barry Street, University of Melbourne. Email: rahsan@unimelb.edu.au.

[‡]Department of Economics, UNSW Business School, University of New South Wales. Email: arpita.chatterjee@unsw.edu.au.

[§]Initiative for Policy Dialogue, Columbia University

# 1 Introduction

Much of the recent literature on trade, inequality, and intergenerational mobility has focused on the effects of international trade, even though the reductions in domestic trade costs through investments in transport infrastructure may be more important, especially in developing countries (Donaldson, 2015). Developing countries spend large sums on transportation infrastructure projects that shape their countries for decades to come. Currently, about 20% of World Bank lending is spent on transportation infrastructure, which is more than the Banks lending for social programs. The literature on transport development tends to find several implications for such development via its effect on agricultural income volatility (Allen and Atkin, 2016), and per capita GDP (Banerjee et al. 2012). However, little is understood of the impact of improvements in domestic transport infrastructure on intergenerational mobility. This paper focuses on a particularly important channel, human capital investment, through which domestic market integration may affect intergenerational mobility in human capital.

Intergenerational mobility is important for both fairness and economic efficiency in a society.[1] A question of long-standing interest among social scientists is the degree to which an individuals status in society is determined by the position of his or her parents. For children born in underpriviledged background, if their outcomes are very tightly linked to that of their parents, this can lead to persistence of inequality despite overall economic growth. Lack of intergenerational mobility is infact a omnipresent sign for lack of inclusive economic development: The Global Database on Intergenerational Mobility (GDIM), compiled by Narayan et al. 2018, convincingly document that developing countries tend to have lower social mobility.[2]

While many outcomes –education, income, occupation and wealth– tend to show intergen-

[1] Brunori, Ferreira, and Peragine (2013) use cross-country data to document a negative correlation between inequality of opportunity and intergenerational mobility.

[2] According to Narayan et al. 2018, "on average, mobility is found to be considerably lower in developing (low- and middle-income) economies when compared to high-income economies (based on the latest World Bank classification). Among developing economies, East Asia and Pacific and Middle East and North Africa are the regions with the highest average relative mobility in education, which are still well below the average for high-income economies...Sub-Saharan Africa and South Asia stand out as regions with some of the lowest levels of mobility. 13 of the 15 least mobile countries are either in Africa or South Asia." Chetty et al. (2014) estimate intergenerational mobility for the US down to highly disaggregated geographic areas and find that lower intergenerational mobility (or higher inequality of opportunity) tends to be concentrated in the lagging areas of the US.

erational persistence, we focus here on persistence in education, because education is widely perceived to be the social ladder to escape povery trap and many adult labor market outcomes, such as income and occupation, crucially depend on education in a market economy. Hertz et al (2007) estimates 50-year trends in the intergenerational persistence of educational attainment for a sample of 42 nations around the globe, and documents large regional differences in educational persistence, with developing countries in Latin America displaying the highest intergenerational correlations. In this paper we ask whether improvements in transport infrastructure and associate domestic market integration leads to lowering this intergenerational persistence thus improving social mobility, or increases this persistence further worsening inequality of opportunity.

The theory of intergenerational mobility, based on Becker etal 2015, does not provide unambiguous predictions; market integration can improve mobility by relaxing binding credit constraint through higher incomes for the poor, or by increasing returns to skilled labor disproportionately for the younger generation. But whether market integration improves income for the poor or returns to education depends on the pattern of specialization, as clearly demonstrated in Krugman and Venables 1995; Baum-Snow et al 2018.[3] Whether market integration improves or worsens intergenerational educational mobility is thus ultimately an empirical question.

We use household-level data from India to empirically examine the relationship between domestic market access and intergenerational education mobility. Our data are drawn from the nationally representative Indian Human Development Survey (IHDS). Unlike other household-level datasets such as the National Sample Survey Organisation data, a key advantage of the IHDS is that all respondents are asked their father's education. Thus, we do not require a father and son to co-reside in the same household for us to examine intergenerational changes in education. This allows us to use a more representative sample of father-son pairs to study intergenerational persistence.[4]

[3]There is evidence that reducing transportation costs can increase (Ghani et al., 2015; Storeygard, 2014), decrease (Faber, 2014) or leave unchanged (Banerjee et al., 2012) growth rates in local economic activity.

[4]Estimates of intergenerational persistence from co-resident samples can have a significant downward bias, which creates a false impression of intergenerational mobility (Emran et al. (2018)). Our sample is not vulnerable to this truncation bias that arises from only including fathers and sons that co-reside in the same household at the time of the survey.

Apart from this data advantage, there are two important reasons why India is an ideal setting in which to examine the relationship between market access and intergenerational education mobility. First, as a geographically large country, India has significant spatial variation in market access. Also, India has experienced dramatic changes in transport infrastructure in recent decades. [5] Second, India is a country where educational attainment is highly persistent across generations. For instance, our data suggest that conditional on having a father who is at the bottom quartile of the fathers' educational distribution in 2012, a son has a 45.50 percent chance of being in the bottom quartile of the sons' educational distribution.[6] Thus, examining role of domestic market integration in lowering intergenerational persistence in education is especially relevant in such a context.

With our household-level data in hand, we examine whether the intergenerational persistence in father-son educational attainment depends on the degree of market access in their district of residence.[7] We follow the conventional approach in the recent economic geography literature and construct a market access measure that captures a district's proximity (by road) to large markets.[8] Thus, districts that are close to a dense road network and/or are close to populous areas will have greater market access.

A key threat to our identification strategy is the non-random placement of road infrastructure. To account for this, we use two instrumental variable (IV) strategies. First, we use historical railroads to instrument the current road network. More precisely, we use the data in Donaldson (2018) to construct an indicator for whether a district had a railway line in 1880 and the length of the railway line in the district in 1880. We then instrument our market access measure using these

[5]See Allen and Atkin (2016) and Ghani, Goswami and Kerr (2016) for evidence of highway constructions post-2000 and Aggarwal (2014) for construction of rural roads. More aggresive push is expected to continue in future with the current Bharatmala initiative under which the fiscal year 2018 experienced highest road construction pace of 27 km per day, almost twice than that of 2014, as documented in National Highway Authority of India website.

[6]Our regression estimates suggest that the intergenerational education correlation in India is 0.535, which is towards the higher end of cross-country estimates reported by Hertz et al. (2017).

[7]We use the two most widely-used measures of intergenerational mobility: intergenerational regression coefficient (henceforth IGRC) and intergenerational correlation (henceforth IGC). For discussions on the interpretation and advantages and disadvantages of these alternative measures of mobility, see Nybom and Stuhler (2017), Emran et al. (2018), Emran and Shilpi (2018).

[8]Our market access measure is based on the approach of Donaldson and Hornbeck (2016) and Allen and Atkin (2016) and uses the bilateral travel time (by road) between Indian districts constructed by the latter.

historical rail infrastructure measures. According to Duranton and Turner (2012), the presence of a historical railway line indicates that an area has a topography that is suitable for transport infrastructure. This should ensure that this same area has more developed road infrastructure in modern times.[9] Further, as Donaldson (2018) points out, the mid to late 19th century railway network in India in was built primarily to facilitate military movement. Thus, the location of this network should be orthogonal to household-level human capital decisions in more recent times.

Our second IV strategy relies on the inconsequential place argument (Redding and Turner (2015)). In particular, we exploit the fact that India's recent highway expansion programs, the Golden Quadrilateral (GQ) and North-South-East-West Corridor (NS-EW), were built to connect the four major urban centers (Delhi, Mumbai, Kolkata, Chennai, and Kolkata). Thus, proximity to these highways provides a village in our sample with greater market access. However, because these highways were not targeted to individual villages, the location of a village with respect to these highways can be considered random.[10] Our instrument for market access in this case is the arc distance between each location in our sample and the nearest point on the GQ or the NS-EW.

Both our baseline OLS and IV results suggest that improved access to markets lowers the correlation between a father and his sons educational attainment. The intergenerational regression coefficient (IGRC) for an individual in a district with a ln(market access) of 0 (i.e. market access of 1) is 0.60. In contrast, the IGRC for an individual in a district with a ln(market access) of 1 (a 2.5 standard deviation increase) is 0.48, which is a substantial reduction in intergenerational persistence. These results suggest that in districts with greater market access, an individual's educational attainment is less tied to the educational attainment of his parents.

We uncover an interesting patterns of heterogeniety in the impact of market access. First, we observe variation in the impact of market access depending on the level of schooling. The estimates from IV quantile regressions using a control function approach deliver a set of interesting conclusions. Better market access improves educational mobility across the entire conditional dis-

[9] Past infrastructure has been used as an instrument for modern infrastructure by, among others, Baum-Snow (2007), Michaels (2008), and Baum-Snow et al. (2017).

[10] Note that we are not using the recent expansion of the GQ and NS-EW in a difference-in-difference design. Much of this highway network had existed in the past, albeit with lower-quality roads. What is important for this IV strategy is the proximity of a village to this highway network, which provided villages with exogenous variation in market access.

tribution of children's schooling, but the strongest effect is felt for the children at the middle of the distribution. At any given level of market access, the children at the lower tail of the distribution face the lowest mobility, and mobility monotonically improves as we move up the conditional distribution.

The contributions of this paper intersect two important and active areas of economic research: (i) intergenerational mobility in developing countries and (ii) economic effects of market integration. For the latter, the existing literature has shown that transportation infrastructure raises the value of agricultural land (Donaldson and Hornbeck, 2015), increase agricultural trade and income (Donaldson, 2018), reduces the risk of famine (Burgess and Donaldson, 2012), increases migration (Morten and Oliveira, 2014), and accelerates urban decentralization (Baum-Snow et al., 2015). This literature largely focuses on quantifying overall gains from trade. For example, Allen and Atkin (2016) quantify welfare gains from trade via effect of domestic market integration on volatility. Morten and Olivera (2017) also quantify welfare effects of improved product and labor market access. However, little is understood of the distributional impact of such integration. By focusing on intergenerational mobility and hence inequality of opportunity, we fill an important gap in this literature.

Our paper also contributes to the literature on intergenerational mobility in developing countries. This literature has focused on measurement and understanding the trends over the last few decades (see, for example, Hertz et al. (2007) and Neidhofer et al. (2018)). In the context of India, the focus has been on the role played by caste, religion, gender, and heterogeneity across different states (see, among others, Azam and Bhatt (2015), Emran and Shilpi (2015)). Ahsan and Chatterjee (2017) analyse the role of international trade liberalization in explaining intergenerational occupational mobility. To the best of our knowledge, there is no study that analyzes the effects of domestic market integration on intergenerational education mobility.

The rest of the paper is organized as follows. In section 2, we lay out the basic structure of the model. We describe the data in section 3, and describe our identification strategy and empirical results in section 4. Section 5 concludes. Understanding the mechanisms by which market access leads to improvements in income to ultimately impact intergenerational mobility and possible sources of heterogeniety form the core of our future work.

## 2 A Model of Intergenerational Educational Mobility

This section presents a simple model of intergenerational mobility in human capital. We follow the approach of Becker et al (2015). We assume that there are two periods of life: childhood and adulthood. Each parent has one child at the beginning of adulthood, which means parents and children overlap when the latter are young. Adults use the human capital accumulated as children to generate labor income. The earnings can be spent on consumption and investment in the human capital of their children, or as bequest.

Parental preferences are assumed to depend not only on parents' consumption, $c$, but also on the expected utility of children. A natural formulation is

$$V(Y_p) \quad = \quad u(c) + \beta \mathbb{E}[V(Y_c)] \tag{1}$$

where $\beta$ is parents' degree of altruism toward their children, and $Y_p$ and $Y_c$ are the monetary resources of parents and children, respectively. We assume that the value function V exists and is continuously differentiable. Parents choose consumption $c$, their human capital investment $z$ and bequest $b_c$. Market rate of interest is $R_k$. The budget constraint for the parents' choice problem is

$$c + z + \frac{b_c}{R_k} = Y_p \tag{2}$$

Parent's choice between consumption, bequest and human capital investment is guided by their altruism and the impact of human capital investment on children's human capital and hence, on income. We assume that for every generation their income is a (weakly) increasing function of their human capital and is also potentially impacted by the market access (MA) prevalent in the period of their adulthood. Thus,

$$Y_i \quad = \quad f(H_i, MA_i), i = p, c \tag{3}$$

We make the standard assumption about the income function that returns to human capital is positive:

$$r_i \equiv \frac{\partial Y_i}{\partial H_i} > 0, i = p, c \tag{4}$$

However, this structure does not impose any assumption on how market access affects income, $\frac{\partial Y_i}{\partial MA_i}$, or returns to human capital, $\frac{\partial r_i}{\partial MA_i} \equiv \frac{\partial^2 Y_i}{\partial H_i \partial MA_I}$. We study the effect of current market access on income and returns to human capital in detail in the next section.

Directly following Becker et al (2015), we assume

$$H_c \quad = \quad \mu + \kappa z + \phi z^2 + \delta H_p \tag{5}$$

Clearly, we expect that $\kappa > 0$ and $\delta > 0$, i.e. that increases in parental human capital and their investment in children's human capital both raise the human capital of children. Also, it is standard to assume that there is diminishing return to $z$, since it becomes harder and harder to instill more knowldge into children with fixed mental capacity, i.e. $\phi < 0$. [11]

Presence of credit constraint for human capital implies parents with higher income face lower borrowing cost, while the richest parents have constant marginal costs of funds. More formally, we assume that the interest rate as a function of earnings, $R(Y_p)$, satisfies

$$\frac{dR}{dY_p} \leqslant 0, \frac{d^2R}{dY_p} \geqslant 0 \tag{6}$$

with $\frac{dR}{dY_p}$ equal to 0 and $R(Y_p)$ equal to $R_k$ for $Y_p$ large enough.[12] We solve for optimal investment

[11] For simplicity, we abstract away from complementarity between parental human capital and their investment in children, other (environmental) influences on children's human capital as well as possibility of nonlinear effects of parents' human capital on children.

[12] The idea behind this modelling of credit constraint is that access to credit for human capital is more restricted among poorer households. There is lot of evidence of such income dependency in access to credit market in India. In rural India, the formal credit market is not very developed. Munshi and Rosenzweig (2009) document that large part of informal loans in rural India are from caste-based network of relatives. They model such informal lending on the basis of mutual insurance and provide empirical evidence that given any household's own income, they are less likely to participate in mutual insurance with poorer network members thereby limiting credit access of those members.Rani (2017) documents phenomenal growth in student loans in India since their introduction in 2001 at an

in human capital both with and without credit constraint.

In presence of credit constraint, the optimal investment in human capital is given by

$$z^* = \frac{\frac{R(Y_p)}{r_c} - \kappa}{2\phi} \tag{7}$$

It is straightforward to show that both investment in human capital as well as children's human capital increases in parental human capital, as shown below:[13]

$$\frac{\partial z^*}{\partial H_p} = \frac{\frac{dR}{dY_p}\frac{r_p}{r_c}}{2\phi} > 0 \tag{10}$$

$$\frac{\mathrm{d}H_c}{\mathrm{d}H_p} = \frac{R(Y_p)}{2r_c\phi}\frac{dR}{dY_p}\frac{r_p}{r_c} + \gamma > 0 \tag{11}$$

The crucial comparative static exercise in our context is how improvement in market access, MA, leads to changes in parental investment in human capital and persistence of human capital. We assume that improvement in market access in adulthood, $MA_p$, can not affect $H_p$ since human capital is formed during childhood. However, change in $MA_p$ can potentially impact $Y_p$. If $\frac{\partial Y_p}{\partial MA_p} > 0$ without any impact on $H_p$, that automatically implies returns to parents' human capital, $r_p$, improves with market access: $\frac{\partial r_p}{\partial MA_p} > 0$. Consider the case when improvement in $MA_p$ both

annual average growth rate of 38 per cent. Using an unique data on education loans made available by Reserve Bank of India, Rani (2017) shows that average interest subsidy and size of loan is larger for students from higher quintiles of parental income which reflects a relatively starker lack of credit market for poorer households. This is reflected in official education loan documents of the largest bank of India, State Bank of India which clarly states: "If the candidate is taking a student loan up to four lakhs, the only security needed is the parent or guardian as the co-borrower. If the loan amount is between 4 lakhs and 7.5 lakhs, the candidate must provide parent or guardian as the co-borrower and collateral security in the form of the details of the Gross Annual Income of the parent or guardian. If the candidate is applying for an education loan above 7.5 lakhs, the parent or guardian must be the co-borrower and the candidate must show tangible collateral security." These requirements for security to take education loans reflect an income and asset dependence in borrower's access to formal credit.

[13]The corresponding expressions without credit constraint are:

$$z^* = \frac{\frac{R_k}{r_c} - \kappa}{2\phi} \tag{8}$$

$$\frac{dH_c}{dH_p} = \gamma > 0 \tag{9}$$

increases $Y_p$ and increases children's return from human capital at a (weakly) faster rate than that of parents such that $\frac{\partial \frac{r_p}{r_c}}{\partial MA_p} \leqslant 0$. In this case, in presence of credit constraints, improvement in market access both increases investment in human capital and reduces persistence of human capital:

$$\frac{\partial z^*}{\partial MA_p} = \frac{1}{2\phi}[\frac{1}{r_c}\frac{dR}{dY_p}\frac{\partial Y_p}{\partial MA_p} + R(Y_p)(-{r_c}^2)\frac{\partial r_c}{\partial MA_p}] > 0 \tag{12}$$

$$\frac{d^2 H_c}{dH_p dMA_p} = \frac{1}{2\phi}\frac{dR}{dY_p}\left[\frac{dR}{dY_p}\frac{\partial Y_p}{\partial MA_p}\frac{r_p}{{r_c}^2} + R(Y_p)\frac{\partial \frac{r_p}{{r_c}^2}}{\partial MA_p}\right] + \frac{d^2 R}{dY_p^2}\frac{\partial Y_p}{\partial MA_p}\frac{R(Y_p)}{2\phi}\frac{r_p}{{r_c}^2} < 0 \tag{13}$$

But in absence of credit constraint, improvement in $MA_p$ has no impact on intergenerational persistence even if $\frac{\partial Y_p}{\partial MA_p} > 0$.[14] Thus, improvement in market access reduces intergenerational persistence in human capital provided such improvement leads to increase in parental income or a disproportionate rise in children's return to human capital.

# 3 Data

## 3.1 Household Data

Our individual-level data are drawn from the Indian Human Development Survey (IHDS). IHDS is a nationally representative individual survey that is available over two rounds: 2004–05 and 2011–12 (Desai et al. (2005); Desai and Vanneman (2012)).[15] To avoid including duplicate observations, we restrict our sample to the 2012 IHDS round. For instance, consider an individual who is 30 years of age in 2005. It is likely that by this point, this individual has completed his formal schooling. As a result, this person's year of schooling will be identical in both survey rounds. Further, the market access measure we construct below is specific to an individual's birth decade. Thus, this time invariant measure will also not change for this individual between the IHDS survey rounds. As a result, including him in both rounds will create a duplicate observation.

[14] From equation (9), $\frac{d^2 H_c}{dH_p dMA_p} = 0$. Note that in absence of credit constraint, improvement in $MA_p$ increases $z^*$ as long as $MA_p$ also improves children's return from human capital ($\frac{\partial r_c}{\partial MA_p} > 0$).

[15] In the remainder of this paper, we refer to the survey year using the second year of the survey. In other words, we refer to the 2011–2012 round as 2012.

To avoid this problem, we only include the latest round of data in our working sample.[16]

The raw data includes both rural and urban areas. Given our goal of estimating the effect of improved market access on educational mobility, we restrict our sample to rural areas. This allows us to focus on households for whom market access is an especially important determinant of income. In addition, we follow the literature (e.g. Hnatkovska et al. (2013)) and exclude females from our analysis to focus on father-son educational correlations. This allows to more readily compare our mobility estimates with the previous literature. We describe the construction of our working sample in greater detail in section A.1 in the Appendix. Our final sample consists of 37,695 observations in 242 rural districts across India.

To measure intergenerational education mobility, we need to first find the father's education for each male in our sample. We do so using the following steps. First, for men that are classified as sons of the household head, we substitute the head's education as the father's education. Unlike other comparable household surveys, e.g. the NSSO household surveys, a key advantage of the IHDS data is that it asks each respondent what their father's educational attainment is. This allows us to find the father's education of the household head themselves even if the father does not reside in the same household or if the father is deceased. In contrast, in standard household surveys such as the NSSO, we can only pair fathers and sons that co-reside in the same household. Further, the IHDS data also allow us to pair multiple father and sons within the same household. For instance, suppose a household consists of a male household head, his brother, son, and nephew. The IHDS data allow us to pair the household head with his son and also pair his brother with his own son (i.e. the head's nephew). These advantages mean that the IHDS data allow us to use a more representative sample of father-son pairs (Azam and Bhatt (2015)).

We report the educational profile of the individuals in our sample in Table 1. These numbers suggest that the average son in our sample has 6.49 years of schooling while the average father in our sample has 2.84 years of schooling. Our data also suggest that, on average, 5.10 percent of these sons have zero years of schooling while a further 41.50 percent have between one and five

[16] In a robustness check, we show that our core result is robust to including just the earlier round of data (IHDS 2005) instead.

years of schooling. Only 18.50 percent of sons in our sample have between 11 and 15 years of schooling. Further, 75.50 percent of sons in our sample are literate, which is defined as someone who can read and write a sentence, and 76.70 percent have ever attended school. Overall, the numbers reported in Table 1 suggest that our sample consists of sons with very moderate levels of formal schooling.

We now turn to the degree of intergenerational educational persistence in our data. To illustrate this, we consider two types of sons. First, we examine sons that have fathers with bottom-quartile schooling attainment among the fathers' educational distribution.[17] We then examine the years-of-schooling distribution of all sons in this sub-sample, which is illustrated in Figure 1. This figure suggests that, conditional on having a bottom-quartile father, a son in our sample has a 45.50 percent chance of being in the bottom quartile of the sons' educational distribution.

Next, we restrict the sample to sons who have fathers with top-quartile schooling attainment among the fathers' educational distribution. We then examine the years-of-schooling distribution of all sons in this sub-sample, which is illustrated in Figure 2. This figure suggests that, conditional on having a top-quartile father, a son in our sample has a 39.36 percent chance of being in the top quartile of the sons' educational distribution. Thus, not only does our sample consist of sons with low levels of schooling attainment, these sons have schooling attainment that is highly correlated with that of their fathers. Summary statistics for other demographic characteristics of the individuals in our sample are provided in Table 2.

### 3.2 Constructing Market Access

To construct our market access variable, which we define below, we require data on bilateral trade costs between districts. Our proxy for bilateral trade costs is the travel time between districts in India, as constructed by Allen and Atkin (2016). They constructed these bilateral travel time using various editions of the *Road Map of India* produced by the Indian government. They first digitized and geo-coded these maps and used an algorithm based on the color of digitized pixels

[17] All distributions used to illustrate Figures 1 and 2 are calculated using the 2012 survey round.

to determine the location of highways. They then used this map to calculate the travel time in hours (by road) between every district in their data by assigning a travel speed of 60 miles per hour on highways and 20 miles per hour on other roads. They constructed these travel time measures for the years 1962, 1969, 1977, 1988, 1996, 2004, and 2011.

We use these travel time data to define individual $i$'s market access ($MA$) as follows

$$MA_{ioc} = \sum_{d \neq o} s_{odc}^{-\phi} N_d \tag{14}$$

where $i$ indexes individuals, $o$ indexes origin districts, $c$ indexes an individual's birth decade, and $d$ indexes destination districts. $s_{odc}$ is the travel time (in hours) between districts $o$ and $d$ in decade $c$. Recall that the Allen and Atkin (2016) travel time data are available between 1962 and 2011. We exploit this time coverage to assign an individual the market access value in his district during his decade of birth. We discuss this choice in more detail below. $\phi$ is a parameter that governs how quickly market access declines with improved travel time, and $N_d$ is the size of the destination district.[18] Thus, this proxy for market access is a function of the density of a district's road network, its proximity to populous neighbors, and the decade-of-birth of an individual. This means that it has both individual-level and district-level variation.

Our choice of constructing market access by birth decade merits further discussion. Recall that our sample consists of men aged 15 years or over in 2012. One option would be to regress an individual's years of schooling in 2012 on the market access of his district of residence in the same year. However, as we showed in section 2 above, market access affects a child's education through its impact on his parents' income during his schooling years. Thus, contemporaneous market access is not the most relevant measure in our context. Further, given that the average years of schooling in our data is 6.48 years, the majority of individuals in our sample would have already completed their formal schooling at a young age. This means that for these individuals, contemporaneous changes in market access should have no effect on their years of schooling. Instead, the market

[18] Following Allen and Atkin (2016), we chose a baseline value of $\phi = 1.5$. Further, we capture the size of the destination market using its population in 2005. This is the first year for which we have population data in our IHDS sample.

access that matters for these individuals is the market access during their school-going age. We approximate this by assigning each individual their district's market access during their decade of birth. That is, for an individual born in the 1960's, we assign him his district's market access in 1969, for an individual born in the 1970's, we assign him his district's market access in 1977 and so forth. For individuals born before the 1960's, we assign them their district's market access in 1962, which is the earliest year for which the travel time data are available.

We illustrate the relationship between market access and son's education in Table 3. Here we place the sons in our sample in to cells depending on their father's education quartile and market-access quartile and then examine there educational attainment. For instance, in the first cell of Table 3, we report the mean years of schooling and its standard deviation in brackets for sons with fathers with bottom-quartile educational attainment and residing in areas with bottom-quartile market access. For example, our data suggest that these sons have 4.29 years of schooling. Two important patterns are immediately clear from the numbers reported in Table 3. First, regardless of the market access quartile, a son's educational attainment is always increasing in the father's educational quartile. Second, for any given father's education quartile, greater market access is associated with greater son's education. However, this association is less clear when we compare sons in the second- and third-quartile market access.

Finally, to implement our instrumental variable (IV) strategy, we exploit two sources of exogeneous variation in market access: (a) historical railroad and (b) distance to the Golden Quadrilateral and the North-South and East-West Corridor. We construct the former using data from Donaldson (2018). More precisely, for each district in our data, we use as instruments an indicator for whether or not that district had a railway line in 1880 as well as the natural logarithm of the length of the railway line in that district in 1880. We use the 1880 railway lines since it preceded the 1880 Famine Commission, which recommended that post-1880 railway lines be targeted towards regions that experienced drought during the late 1870's (Donaldson (2010)). Thus, the railway lines that existed in 1880 were unaffected by the Commission's recommendations. For the Golden Quadrilateral and the North-South and East-West Corridor, we calculate the arc distance between each district in our sample and the nearest point on these respective highway networks.[19]

# 4 Empirical Issues and Strategy

To understand the implications of market access for educational opportunities, we analyze how better market access affects two most widely used measures of intergenerational mobility: intergenerational regression coefficient (IGRC) and intergenerational correlation (IGC). The standard regression specifications for estimating IGRC and IGC can be written as follows (taking into account the role of market access):

$$S_{ij}^{c} = \alpha_1 + \alpha_2 S_{ij}^{f} + \alpha_3 \ln\left(MA_j\right) + \alpha_4 S_{ij}^{f} \times \ln\left(MA_j\right) + \epsilon_{ij} \tag{15}$$

$$\frac{S_{ij}^{c}}{sd^{c}} = \beta_1 + \beta_2 \frac{S_{ij}^{f}}{sd^{f}} + \beta_3 \ln\left(MA_j\right) + \beta_4 \frac{S_{ij}^{f}}{sd^{f}} \times \ln\left(MA_j\right) + \kappa_{ij} \tag{16}$$

where $S$ denotes schooling attainment, $sd$ is standard deviation of schooling, the subscripts $i$ and $j$ refer to an individual (son) and the district in which individual $i$ is located in, respectively; the superscripts $c$ and $f$ refer to the son (children) and father, respectively, and $\ln\left(MA_j\right)$ measures access to markets for district $j$. The specification in equations (15) and (16) above allows better market access to affect both the intercept and the slope of the intergenerational persistence equation. The slope parameters provide estimates of IGRC (equation (15)) and IGC (equation (16). For example, $\alpha_2$and $\beta_2$ provide estimates of IGRC and IGC respectively for the districts with $\ln\left(MA_j\right) = 0$, i.e., $MA_j = 1$.[20] The focus of our analysis is on the parameters $\alpha_3$ , $\alpha_4$, $\beta_3$ and $\beta_4$. If better market access is complementary to (a substitute of) father's education in determining a child's educational attainment, then the sign of the parameters $\alpha_4$ and $\beta_4$ is positive (negative). As noted before, understanding whether markets are substitutes or complements of family background is a central focus of this study. It is important to appreciate that our goal is to provide

[19] We than Brian Blankespoor for providing us with these distance calculations.

[20] The average of $\ln\left(MA_j\right)$ in our data is 1.55 and the minimum and maximum are ?? and ??, respectively.

credible estimates of the effects of better market access on IGRC, and IGC, and thus the issue of causal effects of parental education falls beyond the scope of this paper.

Most of the existing studies on intergenerational mobility in developing countries including India estimate equations (15) and (16) using OLS (with $\alpha_3 = \alpha_4 = \beta_3 = \beta_4 = 0$).[21] The recent literature has identified and underscored a number of sources of bias in the estimates of intergenerational persistence when data come from standard household surveys such as LSMS, and NFHS and NSSO in the context of India. There has been a growing emphasis on understanding the biases that arise from data limitations. Most of the household surveys suffer from truncation of the sample because coresidency is used to define household membership. When the sample consists of only the coresident (at the time of the survey) parents and children, the estimates of standard measures of intergenerational persistence such as intergenerational regression coefficient can be seriously biased downward (Emran et al. (2018)).[22] The data we use are not restricted to only the coresident members of a household, and thus are not subject to the truncation bias due to coresidency restrictions in any significant way. A large literature in the context of intergenerational income persistence in developed countries shows that measurement error can lead to substantial attenuation bias in the estimated intergenerational persistence, leading to false impressions of higher economic mobility across generations (Solon (1992), Mazumder (2005)). While measurement error in income data is likely to be very high in a rural economy with a predominance of non-market activities, measures of schooling attainment are much more reliable (Deaton (1997)). Since we focus on educational mobility, attenuation bias due to measurement error is thus not likely to be a major concern.

An important issue for our empirical analysis is that the estimated effect of a better market access on intergenerational educational mobility is likely to be biased if we rely on the OLS es-

[21] See, for example, Emarn and Shilpi (2015), Azam and Bhatt (2016?), Neidhofer et al. (2018). Among the few papers that estimate causal effects on intergenerational mobility in the context of developing countries, Ahsan and Chatterjee (2017) provide an analysis of the effects of trade liberalization in India on occupational mobility. Emran and Shilpi (2011) provide estimates of causal effects of parental occupation on children's occupation choice in Vietnam and Nepal using the lower bound of Altonji et al. (2005).

[22] While the estimates of slope-based measures such as IGRC are downward biased in coresident samples, the evidence presented in Emran and Shilpi (2018) shows that the estimate of the intercept term is biased upward. An important finding is that the estimates of IGC suffer much less downward bias in the coresident samples.

timator. The bias arises because of the fact that the placement of transport infrastructure such as roads are not random; economic, social, and political factors are important determinants of whether a road is built to connect a certain village to the growing urban centers. It is, however, difficult to pin down the direction of such biases from a priori reasoning alone. To understand the nature of omitted variables bias in our context, consider the empirical model in equation (15) above. The OLS estimates of the parameters of interest are likely to be biased because, in general, $Cov\left(\ln\left(MA_j\right), \epsilon_{ij}\right) \neq 0$. For example, if the overriding goal of infrastructure investment is geographic targeting for poverty reduction and tackling regional inequality, we would expect that roads will be targeted to the poorer villages with adverse economic endowments, making $Cov\left(\ln\left(MA_j\right), \epsilon_{ij}\right) < 0$. In contrast, when the placement of roads is motivated by efficiency and growth, or tax revenue collection objectives, the roads will be built to connect the villages with better economic potential, and we would expect that $Cov\left(\ln\left(MA_j\right), \epsilon_{ij}\right) > 0$. Another potentially important source of bias is heterogeneity across states in terms of policies related to education and credit market. The upshot of the above discussion is that, the direction of bias in OLS estimates of the parameters in equations (15) and (16) above cannot be determined a priori.

A first step to reducing the biases in the estimates is to include indicators of economic potential of different districts. We include average crop yield, rainfall and temperature as controls in the regressions, and employ state fixed effects to account for state-level variations in policies relevant for education. The regressions also include indicators of religious and caste affiliations, and marital status of a son. The religious affiliation is represented by a Muslim dummy, and caste by a dummy for scheduled caste or tribe. There is substantial evidence that Muslims and scheduled caste/tribe in India face adverse economic opportunities, and they are likely to face credit constraint (refer to caste and credit paper). The extended empirical models are:

$$S_{ij}^{c} = \alpha_1 + \alpha_2 S_{ij}^{f} + \alpha_3 \ln\left(MA_j\right) + \alpha_4 S_{ij}^{f} \times \ln\left(MA_j\right) + \Gamma X_{ij} + \theta_k + v_{ij} \tag{17}$$

$$\frac{S_{ij}^{c}}{sd^{c}} = \beta_1 + \beta_2 \frac{S_{ij}^{f}}{sd^{f}} + \beta_3 \ln\left(MA_j\right) + \beta_4 \frac{S_{ij}^{f}}{sd^{f}} \times \ln\left(MA_j\right) + \Pi X_{ij} + \theta_k + \xi_{ij} \tag{18}$$

where $\theta_k$ is the state fixed effect, and the vector $X_{ij}$ represents a vector of other control variables. The inclusion of relevant omitted variables in the empirical model is expected to reduce bias and we expect that $\left|Cov\left(\ln\left(MA_j\right), \epsilon_{ij}\right)\right| > \left|Cov\left(\ln\left(MA_j\right), v_{ij}\right)\right|$. But it is not possible to identify and include all of the relevant omitted variables that can potentially affect the placement of transport infrastructure and the educational outcomes in a district. Thus it is likely that $\left|Cov\left(\ln\left(MA_j\right), v_{ij}\right)\right| \neq 0$, and estimates of the parameters in equations (17) and (18) applying OLS may still be substantially biased.

### 4.1 Two Sources of Identifying Variation, and Instrumental Variables Strategies

To address the bias in the estimates of equations (17) and (18), we develop two identification strategies that exploit different sources of plausibly exogeneous variations in market access across districts. The first approach takes advantage of the railroad construction during the British colonial period, following the influential study by Donaldson (2018). From 1853 to 1930, 67,247 km long railroad network expanded and brought inland districts out of near-autarky by connecting them with the urban centers and seaports. The identification in this case relies on the argument that the placement of the historical railroad by British colonial rulers was primarily determined by military considerations, and thus can be treated as quasi-experimental with respect to economic outcomes in a district (for an extended discussion, see Donaldson (2018)). The exclusion restriction imposed on historical railroad is strengthened by the inclusion of the controls for economic potential of a district as discussed above. The identifying assumption we use is that, conditional on the controls for economic potential of a district, the historical railroad affects the educational outcomes of the children only through the market access.

The second identification strategy relies on the inconsequential place argument (Redding and Turner (2015)) in the context of the golden quadrilateral (GQ) and north-south east-west corridor (NS-EW). The fundamental insight here is that these highways were designed and built to connect major urban centers to each other and the location of a village with respect to these highways can be considered random because the highways were not targeted to the villages. Consider, for example, the GQ which connects the four major urban centers in India: Delhi, Mumbai, Kolkata, and Chennai (see Figure 1). The rural districts under study in this paper are shown in the graph. Whether a district falls close to an arm of the GQ is quasi-random, as it is inconsequential whether the district is economically well-off or not. However, note that we are not exploiting the recent expansion of the GQ and NSEW in a difference-in-difference design.[23] The fact that at least part of the highway network provided important transport links before the recent expansion allows us to analyze decade specific market access, and how changing market access over time and across geographic space affects educational opportunities in an instrumental variables framework.

We develop instrumental variables strategies utilizing the exogeneous variations in market access of different villages across different districts created by historical railroads, and the GQ and NSEW highway network. More specifically, we use an indicator of presence of railroad in a district in 1880, the railroad length in a district in 1880, and arc distance from a district to GQ (henceforth AD-GQ) and to NSEW (henceforth AD-NSEW) as instruments for market access. Following Donaldson (2018), we use the 1880 railway lines since it preceded the 1880 Famine Commission, which recommended that post-1880 railway lines be targeted towards regions that experienced drought during the late 1870's. Thus, the railway lines that existed in 1880 were unaffected by the Commission's recommendations.

It is important to underscore that the access to market experienced by different cohorts in our data set is likely to be different even when they are different generations in the same household located in the same district. This intertemporal variation can arise because of improvements in road quality and increases in speed limits, addition of new nodes to the transport network, or improvements in transport vehicles. This is important in our context because the market access that is relevant for a given child is for the decade when he/she was school-aged. To account for the intertemporal variation in the market access variable, we interact the sources of exogeneous variations with age cohort dummies (decade specific). Taking GQ as an example, the instruments for $\ln(MA)$ in equations (17) and (18) thus include both $\ln(AD-GQ)$ and its interaction with

[23] A potential objection to such a DID design is that a large part of GQ and NS-EW were originally built a long time ago, and there is no clean baseline without exposure to the treatment.

the birth decade dummies. To ensure that the interaction instrument does not capture other time varying factors relevant for education, we directly control for the cohort dummies in all of the IV regressions.[24] We interact the two basic instruments with father's education to generate instruments for the interaction terms $S_{ij}^{f} \times \ln\left(MA_j\right)$ and $\frac{S_{ij}^{f}}{sd^{f}} \times \ln\left(MA_j\right)$; taking GQ as an example, the instruments are $S_{ij}^{f} \times \ln\left(AD - GQ\right)$ and $\frac{S_{ij}^{f}}{sd^{f}} \times \ln\left(AD - GQ\right)$ for equations (17) and (18) respectively.

### 4.2 Beyond the Mean Effects: Quantile Regression Models

The literature on intergenerational economic mobility, both in developed and developing countries, almost exclusively focuses on the mean effects using OLS regressions. This is largely motivated by the advantages of an easily-understood summary measure of mobility that can be useful for policy analysis and policy decisions by bureaucrats and politicians. It is, however, widely recognized that the mean regressions may miss a lot of interesting and policy-relevant heterogeneity (for a discussion, see De Grawe (2004)). To understand potential heterogeneity in the effects of market access across the quantiles of children's education, we use quantile regressions to estimate the empirical models. For IV quantile regressions, we implement the control function approach developed by Lee (2007, 2004).[25] Following the applications of the method in Lee (2004), we estimate the first stage using a median regression, and retrieve the residual. Since there are two endogenous variables, we get two estimated residuals from this procedure. In the second stage, we include a third degree polynomial of each residual as control function terms and estimate using quantile regression.[26]

[24] If we add quadratic age controls for a son in the specification, the main conclusions of the paper remain intact.

[25] Among the few studies of intergenerational mobility that use quantile regression, see De Grawe (2004) on USA, ??? and Kishan (2018) on India. We are not aware of any other study that uses IV quantile method to analyze intergenerational economic mobility.

[26] We note that the estimates and the main conclusions of this paper are robust if we use the mean regression in the first stage instead of the median regression.

# 5 Empirical Results

## 5.1 The Mean Effects

We begin the discussion with the estimates from the mean regressions, using OLS and 2SLS estimators. Table 4 presents the OLS estimates of IGRC and IGC from specifications that do not include market access variables. This specification is used in most of the available studies on intergenerational educational mobility in India and provides us a benchmark. The odd columns in Table 4 report the estimates without any control variables, and the even columns with the set of control variables discussed earlier. The estimate of IGRC in column 1 is 0.56 which is close to the other available estimates for sons in India; for example, 0.63 in Azam and Bhatt (2016) using IHDS 2005 data. Similarly, the estimated IGC in column 3 of Table 4 is 0.54 which tallies well with the other existing estimates; for example, 0.52 using IHDS 2005 by Azam and Bhatt (2016), and 0.55 using 1993 NFHS data by Emran and Shilpi (2015).[27] The estimates in the even columns are substantially lower which suggests that interstate differences in endowment and policy are important in determining the educational opportunities faced by children.

The OLS results for the effects of market access are reported in table 5. The odd columns report estimates from a simple specification where better market access affects the average schooling in a district (i.e, an intercept effect), but does not affect the slope estimates (IGRC and IGC), while the even columns are based on the specifications in equations (17) and (18) above. The evidence from the odd columns shows that a better market access has a positive and statistically significant effect on son's educational attainment which can be interpreted as a "rising tide lifts all boats" effect because it does not depend on the family background of a child as captured by father's education. The estimate in column 1 of Table 5 suggests that the sons born and raised in a district with market access value of 2.7 (i.e., ln(MA)=1) would have 0.27 years of more schooling when compared to the children in a district with market access value of 1 (i.e., ln(MA)=0). The mean value of log market access in our data is 1.55 which implies that the children located at these districts enjoy

[27] It is interesting that the estimate of IGC from NFHS data is so close to the estimates from IHDS data, because NFHS data include only the coresident children in the sample. This is, however, consistent with the finding by Emran et al. (2018) that the truncation bias due to coresidency is low in the IGC estimates.

0.42 years of additional schooling compared to the children in districts with ln(MA)=0. These are substantial effects.

Perhaps, more interesting is the evidence on the effects of better market access on the slope parameters (IGRC in column 2 and IGC in column 4); the estimates show negative and statistically significant (at the 1 percent level) effects for both IGRC and IGC, suggesting that better market access is a substitute for father's education in determining a son's educational attainment. This is consistent with the theoretical predictions in section (3) above, where better market access reduces intergenerational persistence by relaxing credit constraints. To have a sense of the magnitudes involved, consider the estimates of IGRC in column 2 of Table 5; the IGRC for children in districts with a log market access value of 0 (i.e., MA=1) is 0.60 which is a bit larger than the estimate of IGRC= 0.56 in column 1 of Table 4 without any controls. For the sons located in districts with a value of log market access of 1 (i.e., MA=2.7), the IGRC is 0.48, a substantial difference in magnitudes.

### The IV Estimates

The estimates presented in Tables 4 and 5 using the OLS estimator provide interesting evidence that suggests two tentative conclusions: (1) a better market access increases the average schooling attainment in a district, and (2) market access and father's education are substitutes in determining a son's educational attainment. However, as discussed in some detail in section (4) above, the effects of market access estimated by OLS are likely to suffer from substantial bias. In this section we discuss estimates from the instrumental variables strategies developed earlier in section (4) above. The results are reported in Table 6. The first two columns use instruments based on historical railroads; more specifically, the instrument for column 1 is an indicator for whether a district had a rail line in 1880, and in column 2, the instrument is log of the rail line length in a district in 1880. The last two columns in Table 6 present the estimates corresponding to the instruments (i) arch distance between a district and the GQ (column 3), and (ii) arch distance from a district to NSEW highway (column 4).

The first stage F statistics are reported in the lower panel of Table 6. They show that the instruments based on 1880 rail road have reasonable explanatory power for both the endogeneous variables. But the instruments based on arc distance to GQ and NSEW are strong only for the interaction variable, the F statistics are low for first stage regressions for ln(MA) variable. This suggests that the estimates in the first two columns are likely to be more credible. The following discussion on the IV estimates thus focus on the results in the first two columns of Table 6.

There are a number of interesting things to note about the results in Table 6. First, the numerical magnitudes of all the relevant parameters are larger when we instrument for market access and its interaction with father's education. Second, the basic qualitative conclusions from Table 5 based on OLS estimates remain intact: a better market access improves educational attainment for all children (a positive intercept effect), and the interaction effect is negative implying that market access acts as a substitute for family background (father's education).[28] The fact that the IV estimates are larger than the corresponding OLS estimates is consistent with the literature, and suggests that the placement of roads and rail roads are primarily motivated by poverty alleviation and regional inequality. It is also likely that the market access measure constructed from the Allen and Atkin data suffer from significant measurement error, leading to substantial attenuation bias in the OLS estimates of the effects of market access.

To understand better the economic magnitudes involved, consider, for example, the estimates in column 2 (with the highest first stage F statistic). The estimates imply that children in a district with the average log market access have on average almost 3 years of additional schooling when compared to the children located in a district with ln(MA)=0. This is clearly a large effect. The estimate of the interaction term implies that the IGRC for the children in a district with the average log market access (i.e., ln(MA)=1.55) is 0.42; in contrast, the IGRC for the children in districts with ln(MA)=0 is 0.88. Thus the intergenerational persistence in schooling is more than twice as large in a district with low market access (ln(MA)=0). This is striking evidence that transport infrastructure investments have substantially improved intergenerational education mobility in rural India.

[28]It is important to emphasize here that the main qualitative conclusions of this paper remain valid even when the instrument for ln(MA) is weak.

## 5.2 Estimates from the IV Quantile Regressions

The evidence presented in Tables 4-6 and discussed in subsection (5.1) above deals with the mean effects. The focus on the role played by market access and its interaction with father's education provide us a much richer picture of the intergenerational educational persistence in rural India. In this section, we discuss the estimates from IV quantile estimator applied to the empirical models in equations (16) and (17) above. As noted earlier, following Lee (2004), we use a third degree polynomial as the control function in the quantile regression models, but the main conclusions are robust to alternative specifications of the polynomial function (the results are available from the authors).

The estimates from the control function IV quantile estimator are reported in Table 7. We report the estimates for the median, 25th percentile, and 75th percentile. The first three columns contain the estimates when 1880 rail road indicator is the source of identifying variation, and the last three for the case when arc distance to GQ provides the basis for identification. The broad pattern of the estimates is similar across alternative identification schemes. For example, the direct effect of father's education seems to be strongest at the 25th quantile, and the effect weakens monotonically as we move up the conditional distribution. However, the numerical magnitudes vary substantially, and given that the rail road indicator is much stronger in explaining the variations in the market access in the first stage, we confine the following discussion to the results in the first three columns.

The estimates in the first three columns deliver the following conclusions. First, the effect of market access on the average schooling in a district is positive and statistically significant across the distribution, and the highest impact is observed for the median group. Perhaps more important is the evidence that the sons in the 25th quantile gain the lowest in terms of average schooling from a given improvement in market access. Second, the effects of better market access on the estimated IGRC is consistently negative across the distribution, with the highest impact again observed for the median group, and the lowest for the 75th percentile. These estimates suggest that the mean effects in Table 5 miss a substantial amount of heterogeneity across the children's distribution.

To have a better sense of the magnitudes involved, we again consider the districts with ln(MA)=0 versus the districts with the mean level of log access to market (i.e., ln(MA)=1.55). The estimates imply that the average schooling goes up by 0.92 years in a district with mean level of market access (log) for the sons who belong to the 25th percentile of the conditional schooling distribution; the corresponding numbers for the median and 75th percentiles are 3.75 years, and 2.37 years. The gap between the students at the lower part and the top thus widens with better access to markets. For the 25th percentile, the estimate of IGRC is 0.95 if a child is located in a district with ln(MA)=0, but it falls to 0.62 if he is located in a district with the average market access (log). The corresponding IGRC estimates for the median are: 0.83 (ln(MA)=0) and 0.41 (ln(MA)=1.55), and for the 75th percentile are: 0.51 (ln(MA)=0) and 0.32 (ln(MA)=1.55). The evidence thus leads to two important conclusions: (i) better market access reduces the IGRC, but (2) at a given level of market access, the IGRC is the highest for the sons at the lower part of the distribution, and the magnitude of the IGRC declines monotonically as we climb up the conditional schooling distribution of the sons.

To understand the implications of the above findings, it is instructive to consider the interpretation of the error term in regression specifications (17) and (18). As noted by De Grawe (2004), a plausible interpretation of the error term is that it captures innate ability of a child. In this perspective, the estimates imply that the sons with above average ability experience much better educational mobility, especially when compared to the sons at the lowest part of the ability distribution. This may be result of both optimal investment choices by poor credit constrained parents with multiple school-aged children, and the educational policies pursued by government. When considering allocation of limited resources for educational investment, it may be optimal for a parent to skew the investment in favor of the most promising child. Government policies such as free primary schooling and midday meals are expected to relax the binding credit constraint faced by the poor households, and help the children with low ability to complete primary schooling. The evidence that the sons at the 25th percentile face the highest intergenerational persistence in education suggest that such policies have not been successful in helping the low-performing students. The merit-based scholarships provided by the government may reinforce the effects of skewed parental investment effect, making the intergenerational persistence low for the exceptionally gifted children. Rani (2014) provides evidence that children from well-off families exit publicly provided low quality school education to capture the freely provided or highly subsidised high cost and high quality public higher education.

## 6 Conclusion

In this paper we provide causal estimates of impact of market access on children's educational outcome and intergenerational persistence of education. There is robust evidence that improvement in market access leads to higher educational attainment and lower intergenerational persistence thereby improving intergenerational mobility in education. Using quantile regressions, we show that this effect is strongest in the middle of the education disribution. In a standard model of human capital incorporating credit constraint, we show that such effect of market access can only happen if market access improves parent's income and/or children's return to education. Understanding the mechanisms by which market access leads to improvements in income or returns to education to ultimately impact intergenerational mobility and possible sources of heterogeniety in this impact form the core of our future work.

## References

[1] Aggarwal, Shilpa. "Do rural roads create pathways out of poverty? Evidence from India." Journal of Development Economics 133 (2013): 375-395.

[2] Ahsan, Reshad N., and Arpita Chatterjee. "Trade liberalization and intergenerational occupational mobility in urban India." Journal of International Economics 109 (2017): 138-152.

[3] Allen, T., and Atkin, A., 2016. "Volatility and the Gains from Trade." NBER Working Paper No. 22276.

[4] Azam, M., and Bhatt, V., 2015. “Like Father, Like Son? Intergenerational Educational Mobility in India.”*Demography*, 52(6): 1929–1959.

[5] Banerjee, Abhijit, Esther Duflo, and Nancy Qian. On the road: Access to transportation infrastructure and economic growth in China. No. w17897. National Bureau of Economic Research, 2012.

[6] Baum-Snow, Nathaniel. ”Did highways cause suburbanization?.” The Quarterly Journal of Economics 122.2 (2007): 775-805.

[7] Baum-Snow, Nathaniel, et al. ”Roads, railroads, and decentralization of Chinese cities.” Review of Economics and Statistics 99.3 (2017): 435-448.

[8] Brunori, P., Ferreira, G.H.G, and Peragine, V., 2013. “Inequality of Opportunity, Income Inequality and Economic Mobility: Some International Comparisons.” *World Bank Policy Research Working Paper 6304*.

[9] Burgess, Robin, and Dave Donaldson. ”Can openness mitigate the effects of weather shocks? Evidence from India’s famine era.” American Economic Review 100.2 (2010): 449-53.

[10] Chetty, R, N Hendren, P Kline, and E Saez (2014), Where is the land of opportunity? The geography of intergenerational mobility in the United States, Quarterly Journal of Economics 129(4:, 1553-1623.

[11] Desai, S., and Vanneman, R., National Council of Applied Economic Research, New Delhi, 2005. India Human Development Survey (IHDS), 2005. ICPSR22626-v11. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

[12] Desai, S., and Vanneman, R., 2012. India Human Development Survey-II (IHDS-II), 2011-12. ICPSR36151-v5. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

[13] Donaldson, D., 2010. “Railroads of the Raj: Estimating the Impact of Transportation Infrastructure.” Asia Research Centre Working Paper 41, London School of Economics.

[14] Donaldson, D., 2018. “Railroads of the Raj: Estimating the Impact of Transportation Infrastructure.” *American Economic Review*, 108(4–5): 899–934.

[15] Donaldson, Dave, and Richard Hornbeck. ”Railroads and American economic growth: A market access approach.” The Quarterly Journal of Economics 131.2 (2016): 799-858.

[16] Duranton, Gilles, and Matthew A. Turner. ”Urban growth and transportation.” Review of Economic Studies 79.4 (2012): 1407-1440.

[17] Emran, M.S., Greene, W., and Shilpi, F., 2018. “When Measure Matters: Coresidency, Truncation Bias, and Intergenerational Mobility in Developing Countries.” *Journal of Human Resources*, 53(3): 589–607.

[18] Emran, M.S., and Shilpi, F., 2015. “Gender, Geography, and Generations: Intergenerational Educational Mobility in Post-Reform India.” *World Development*, 72: 362–380

[19] Faber, Benjamin, ”Trade Integration, Market Size, And Industrialization: Evidence From China’s National Trunk Highway System,” Review of Economic Studies, 2014, 81.

[20] Ghani, Ejaz, Arti Grover Goswami, and William R. Kerr, ”Highway to Success: The Impact of the Golden Quadrilateral Project for the Location and Performance of Indian Manufacturing,” Economic Journal, 2015.

[21] Grawe, N.D., 2004. “Reconsidering the Use of Nonlinearities in Intergenerational Earnings Mobility as a Test for Credit Constraints.” *Journal of Human Resources*, 39(3): 813–827.

[22] Hnatkovska, V., Lahiri, A., and Paul, S.B., 2013. “Breaking the Caste Barrier: Intergenerational Mobility in India.” *Journal of Human Resources*, 48(2): 435–473.

[23] Hertz, T., et al. ”Intergenerational economic mobility around the world.” The BE Journal of Economic Analysis  Policy 7.2 (2007): 1-46.

[24] Krugman, Paul, and Anthony J. Venables. ”Globalization and the Inequality of Nations.” The quarterly journal of economics 110.4 (1995): 857-880.

[25] Lee, S., 2004. “Endogeneity in Quantile Regression Models: A Control Function Approach.” CeMMAP Working Papers CWP08/04, Institute for Fiscal Studies.

[26] Lee, S., 2007. “Endogeneity in Quantile Regression Models: A Control Function Approach.” *Journal of Econometrics*, 141(2): 1131–1158.

[27] Munshi, Kaivan, and Mark Rosenzweig. Why is mobility in India so low? Social insurance, inequality, and growth. No. w14850. National Bureau of Economic Research, 2009.

[28] Morten, Melanie and Jaqueline Oliveira, , roads and labor market integration: Evidence from a planned capital city,” 2014.

[29] Neidhofer, G, Serrano, J., and Gasparini, L., 2018. “Educational Inequality and Intergenerational Mobility in Latin America: A New Database.” *Journal of Development Economics*, 134: 329–349.

[30] Narayan, A, R Van der Weide, A Cojocaru, C Lakner, S Redaelli, D Mahler, R Ramasubbaiah, and S Thewissen (2018), Fair Progress?: Economic Mobility Across Generations Around the World, Equity and Development, Washington DC: World Bank.

[31] Rani, P. Geetha. ”Financing higher education and education loans in India: interstate differentials and determinants.” Journal of Social and Economic Development 19.1 (2017): 42-59.

[32] Rani, P. Geetha (2014), ”Equity in the Distribution of Indias Government Subsidies on Education” International Journal of Education and Economic Development, 2014, vol.5, pp.1-39.

[33] Storeygard, Adam, ”Farther on down the road: transport costs, trade and urban growth in sub-Saharan Africa,” 2014.

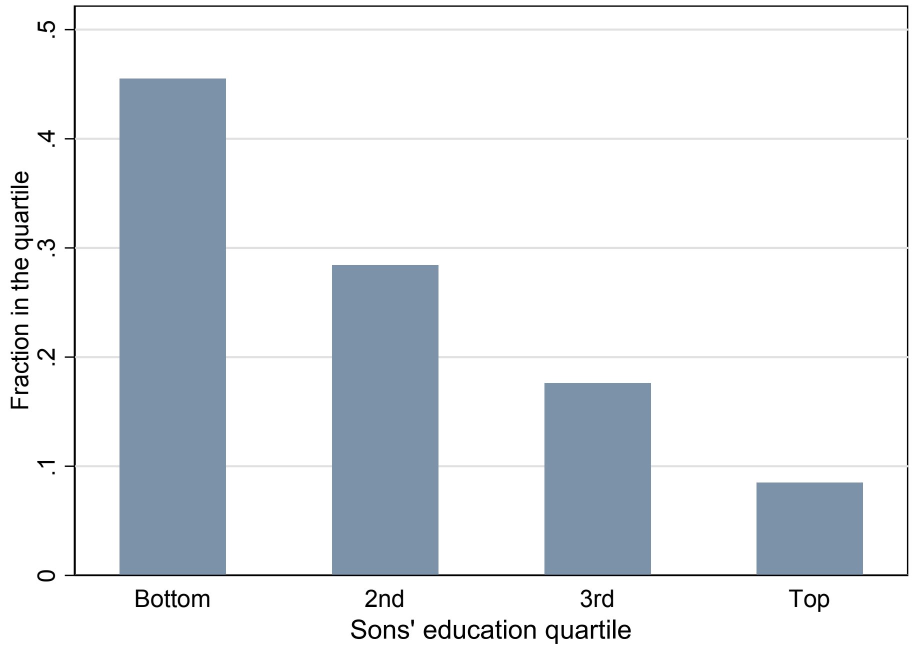


Figure 1: Education Quartiles of Sons Born to Bottom-Quartile Fathers

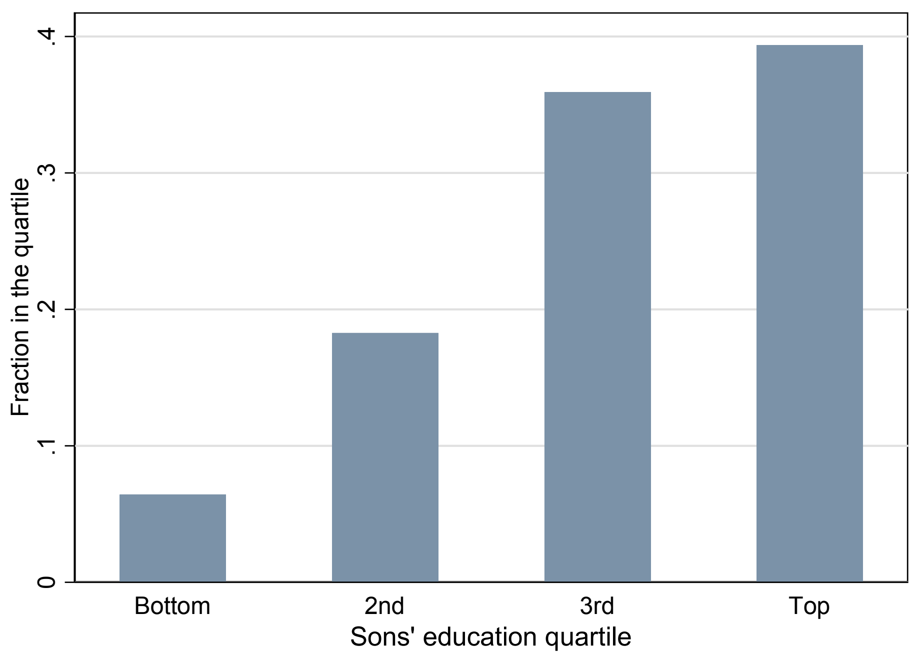


Figure 2: Education Quartiles of Sons Born to Top-Quartile Fathers

Table 1: Education Profile of Sample Individuals

| | (1) |
|---|---|
| Son's Education | 6.483 |
| | [4.644] |
| Fraction Literate | 0.755 |
| Fraction that Ever Attended School | 0.767 |
| Fraction of Sons with | |
| 0 Years of Schooling | 0.051 |
| 1–5 Years of Schooling | 0.415 |
| 6–10 Years of Schooling | 0.399 |
| 11–15 Years of Schooling | 0.185 |
| Father's Education | 2.840 |
| | [3.931] |

**Notes:** son's education is defined as the number of years of schooling completed by each individual in our sample. Father's education is defined as the number of years of schooling completed by each individual's father. For both of these variables, we report the sample mean and standard deviation (in square brackets) above. The remaining variables are categorical, which is why we report fractions. A literate person is defined as someone who reports that they can read and write a sentence.

Table 2: Summary Statistics

| | (1) |
|---|---|
| Son's Education | 6.483 |
| | [4.644] |
| Father's Education | 2.840 |
| | [3.931] |
| Age | 38.705 |
| | [16.848] |
| Married | 0.743 |
| | [0.437] |
| Scheduled Caste/Tribe | 0.323 |
| | [0.468] |
| Muslim | 0.091 |
| | [0.287] |
| Number of Household Members | 5.804 |
| | [2.842] |
| Ln(Market Access) | 1.551 |
| | [0.392] |
| 1880 Railroad Indicator | 0.349 |
| | [0.477] |
| Ln(1880 Railroad Length) | 3.829 |
| | [5.258] |
| Ln(Distance to GQ) | 4.655 |
| | [1.359] |
| Ln(Distance to NSEW) | 4.526 |
| | [1.209] |
| Observations | 37,695 |

**Notes:** son's education is defined as the number of years of schooling completed by each individual in our sample. Father's education is defined as the number of years of schooling completed by each individual's father. The remaining demographic variables are for the sons in our sample. 1880 Railroad Indicator takes the value of one if a district in the sample had a rail line in 1880 while 1880 Railroad Length is the rail line length in a district in 1880. Distance to GQ is the distance between each district in our sample and the nearest point on the Golden Quadrilateral (GQ) highway while Distance to NSEW is the distance between each district in our sample and the nearest point on the NorthSouth and EastWest Corridor highway.

Table 3: Mean and Variance of Son's Education by Father's Education Quartile and Market Access Quartile

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Father's Education Quartile | | | |
| | First | Second | Third | Fourth |
| First Quartile Market Access | 4.286 | 4.504 | 6.467 | 9.438 |
| | [4.272] | [4.491] | [4.360] | [3.851] |
| Second Quartile Market Access | 5.102 | 5.630 | 7.442 | 9.851 |
| | [4.353] | [4.589] | [4.180] | [3.616] |
| Third Quartile Market Access | 4.828 | 5.290 | 7.287 | 10.017 |
| | [4.387] | [4.582] | [4.381] | [3.651] |
| Fourth Quartile Market Access | 5.829 | 6.434 | 8.331 | 10.410 |
| | [4.351] | [4.441] | [3.952] | [3.287] |

**Notes:** to generate the numbers in this table, we first place the sons in our sample into cells based on their father's education quartile and the market access quartile of their district of residence. We then report the mean and standard deviation (in square brackets) of the years of schooling in each cell. For example, the top-left cell above suggests that sons with fathers with the lowest-quartile years of schooling and who also live in districts with the lowest-quartile level of market access have 4.29 years of schooling on average.

Table 4: Intergenerational Education Mobility

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Dependent Variable | Son's Education | | Son's Adjusted Education | |
| Father's Education | 0.562*** | 0.406*** | | |
| | (0.008) | (0.009) | | |
| Father's Adjusted Education | | | 0.535*** | 0.387*** |
| | | | (0.008) | (0.009) |
| Constant | 4.887*** | 8.257*** | 1.033*** | 1.745*** |
| | (0.080) | (0.222) | (0.017) | (0.047) |
| Controls Included? | No | Yes | No | Yes |
| Observations | 37,695 | 37,694 | 37,695 | 37,694 |
| R-squared | 0.226 | 0.329 | 0.226 | 0.329 |

**Notes:** the dependent variable in columns (1) and (2) is the years of schooling attained by a son in our sample. The dependent variable in columns (3) and (4) is a son's years of schooling divided by the standard deviation of all sons' schooling in a given year. The regressions in columns (2) and (4) control for a son's age, age squared, his marital status, an indicator for whether he belongs to a scheduled caste or tribe, and an indicator for whether he is Muslim. The regressions in columns (2) and (4) also include state fixed effects. The standard errors in parenthesis are robust and clustered at the district-year level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 5: Market Access and Intergenerational Education Mobility

| | (1) | (2) | (3) |
|---|---|---|---|
| Dependent Variable | Son's Education | | Son's Adj. Education |
| Father's Education | 0.382*** | 0.605*** | |
| | (0.009) | (0.039) | |
| Ln(Market Access) | 0.269* | 0.605*** | 0.128*** |
| | (0.155) | (0.178) | (0.038) |
| Father's Education × ln(Market Access) | | -0.122*** | |
| | | (0.023) | |
| Father's Adjusted Education | | | 0.576*** |
| | | | (0.037) |
| Father's Adjusted Education × ln(Market Access) | | | -0.117*** |
| | | | (0.022) |
| Observations | 37,694 | 37,694 | 37,694 |
| R-squared | 0.330 | 0.331 | 0.331 |

**Notes:** the dependent variable in columns (1) to (2) is the years of schooling attained by a son in our sample. The dependent variable in column (3) is a son's years of schooling divided by the standard deviation of all sons' schooling in a given year. All regressions include controls for a son's age, age squared, his marital status, an indicator for whether he belongs to a scheduled caste or tribe, an indicator for whether he is Muslim, state fixed effects, and a constant that is not reported. The standard errors in parenthesis are robust and clustered at the district-year level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 6: Endogeneity of Market Access - IV Regressions

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Dependent Variable | Son's Education | | | |
| Instrument | 1880 Railroad Indicator | Ln(1880 Railroad Length) | Ln(Distance to GQ) | Ln(Distance to NSEW) |
| Ln(Market Access) | 1.848*** | 1.975*** | 3.226*** | 1.387 |
| | (0.657) | (0.667) | (0.780) | (0.876) |
| Father's Education | 0.841*** | 0.875*** | 1.448*** | 1.500*** |
| | (0.107) | (0.114) | (0.145) | (0.143) |
| Father's Education $\times$ ln(Market Access) | -0.270*** | -0.291*** | -0.644*** | -0.674*** |
| | (0.064) | (0.068) | (0.086) | (0.085) |
| $F$ stat. for ln(Market Access) | 8.72 | 8.91 | 3.96 | 5.18 |
| $F$ stat. for Father's Education $\times$ ln(Market Access) | 20.97 | 18.91 | 38.34 | 32.67 |
| Observations | 37,694 | 37,694 | 37,694 | 37,694 |
| R-squared | 0.325 | 0.324 | 0.296 | 0.295 |

**Notes:** the dependent variable in all columns is the years of schooling attained by a son in our sample. Let $Z$ be the instrument in any given column above. The full set of instruments in each regression is: $Z$ interacted with birth-decade fixed effects, $Z$ interacted with birth-decade fixed effects and father's education, and father's education interacted with $Z$. All regressions include controls for a son's marital status, an indicator for whether he belongs to a scheduled caste or tribe, an indicator for whether he is Muslim, his decade-of-birth fixed effects, and a constant that is not reported. All regressions also include a district's monsoon rainfall (in natural logarithm), monsoon temperature (in natural logarithm), and its average crop yield (in natural logarithm), and state fixed effects. The standard errors in parenthesis are robust and clustered at the district-year level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 7: Quantile Regressions

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Dependent Variable | Son's Education | | | | | |
| Instrument | 1880 Railroad Indicator | | | Ln(Distance to GQ) | | |
| Quantile | 0.25 | 0.50 | 0.75 | 0.25 | 0.50 | 0.75 |
| Father's Education | 0.952*** | 0.825*** | 0.507*** | 2.074*** | 1.638*** | 1.118*** |
| | (0.051) | (0.054) | (0.057) | (0.068) | (0.057) | (0.054) |
| Ln(Market Access) | 0.594*** | 2.418*** | 1.528*** | 1.152*** | 3.909*** | 2.981*** |
| | (0.161) | (0.321) | (0.381) | (0.208) | (0.366) | (0.609) |
| Father's Education | -0.214*** | -0.270*** | -0.122*** | -0.906*** | -0.761*** | -0.492*** |
| $\times$ ln(Market Access) | (0.031) | (0.032) | (0.034) | (0.039) | (0.033) | (0.033) |
| Constant | -1.689** | -2.726** | 4.428*** | -2.056** | -3.740*** | 2.560** |
| | (0.710) | (1.100) | (0.918) | (0.861) | (0.892) | (1.044) |
| Baseline Schooling | 2 | 7 | 10 | 2 | 7 | 10 |
| Observations | 37,694 | 37,694 | 37,694 | 37,694 | 37,694 | 37,694 |

**Notes:** the dependent variable in all columns is the years of schooling attained by a son in our sample. All regressions include controls for a son's marital status, an indicator for whether he belongs to a scheduled caste or tribe, an indicator for whether he is Muslim,and his decade-of-birth fixed effects. All regressions also include a district's monsoon rainfall (in natural logarithm), monsoon temperature (in natural logarithm), and its average crop yield (in natural logarithm), and state fixed effects. The standard errors in parenthesis are bootstrapped with 50 repetitions. *** p<0.01, ** p<0.05, * p<0.1

# A Data Appendix

## A.1 Construction of the IHDS Working Sample

Our working sample consists of males above the age of 15. Omitting men below 15 years of age ensures that our sample consists of men who have completed their education.[29] This ensures that their years of schooling is not subject to any contemporaneous shocks. We also restrict the sample to men that are directly related to each other. That is, we omit all son-in-laws, other relatives, and domestic servants within a household. This is necessary since we need to match each individual in our sample with his biological father. Such matching is not possible for men who are not directly related to other household members. Lastly, we remove father-son pairs with either missing location or education data.

## A.2 Merging the Allen and Atkin (2016) Data with IHDS

The Allen and Atkin (2016) data do not cover all districts in our IHDS working sample. There are two reasons for this. First, their data do not include smaller states and union territories. As a result, we were unable to measure the travel time to and from these districts in our sample. Second, their data span the years 1962, 1969, 1977, 1988, 1996, 2004, and 2011. To ensure consistency, they use 1966 district boundaries. This means that any districts (and states) that were created after 1966 were not included separately in their data. As a result, these districts appear in the IHDS data but not in the Allen and Atkin (2016) data. To account for this, we create a concordance between each newly created district in our IHDS data and the district that is was a part of in 1966.[30] With the help of this concordance table, we were able to match 330 of the 371 districts in our working sample to the Allen and Atkin (2016) data. These 330 districts represent 92.30 percent of the population in our working sample.

[29] As we showed above, the average individual in our sample has completed approximately 6 years of schooling.

[30] As an example, the Dhubri district in Assam appears in our IHDS data but not in the Allen and Atkin (2016) data. This is because Dhubri was created on July 1st, 1983 when it was split from Goalpara district in Assam. Using our concordance, we replace Dhubri with Goalpara before merging our IHDS data with the Allen and Atkin (2016).

# B Additional Results

Table B.1: Correlation between Market Access and the Instruments

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | Ln(Market Access) | 1880 Rail Indicator | Ln(1880 Rail Length) | Ln(Dist. to GQ) | Ln(Dist. to NSEW) |
| Ln(Market Access) | 1.000 | | | | |
| 1880 Railroad Indicator | 0.302 | 1.000 | | | |
| Ln(1880 Railroad Length) | 0.292 | 0.994 | 1.000 | | |
| Ln(Distance to GQ) | -0.157 | -0.143 | -0.143 | 1.000 | |
| Ln(Distance to NSEW) | -0.260 | -0.160 | -0.137 | -0.142 | 1.000 |