# Estimating Intergenerational Mobility with Coarse Data: A Nonparametric Approach*

Sam Asher†

Paul Novosad‡

Charlie Rafkin§

September 2017
First version: February 2016

**Abstract**

Studies of intergenerational mobility often focus on coarsely measured variables such as education and occupation. Because parent-child correlations within bins are unobservable, conventional rank mobility measures can be biased or uninformative when cells are large. We develop a nonparametric method to bound many measures of mobility, and we propose a new measure that is robust to coarse data. Applying the method to India, we show that recent mobility gains of low caste groups have been offset by significant losses among Muslims, a result that traditional methods miss. Our method has broad applicability, including to current debates on education and mortality.

*We are thankful for useful discussions with Emily Blanchard, Raj Chetty, Eric Edmonds, Shahe Emran, Francisco Ferreira, Nate Hilger, David Laibson, Ethan Ligon, Erzo Luttmer, Nina Pavcnik, Bruce Sacerdote, Na'ama Shenhav, Forhad Shilpi, Gary Solon, Doug Staiger and Chris Snyder. We also thank seminar participants at the Aix-Marseilles School of Economics, the Barcelona GSE Summer Forum and the World Bank Development Research Group for helpful comments, and participants at the Dartmouth College Junior Seminar for asking the questions that led to this research. Taewan Roh provided excellent research assistance. We are grateful to Bratberg et al. (2017) for sharing data. All errors are our own.

†World Bank, sasher@worldbank.org

‡Dartmouth College, paul.novosad@dartmouth.edu, corresponding author

§National Bureau of Economic Research, crafkin@nber.org

## I Introduction

The intergenerational transmission of economic status, a proxy for equality of opportunity, has implications for inequality, allocative efficiency and subjective welfare (Solon, 1999; Black and Devereux, 2011). Studies of intergenerational mobility typically rely upon some measure of rank in the social hierarchy which can be observed for both parents and children (Chetty et al., 2014a; Chetty et al., 2017b).

When the measure of social status is observed only in coarse bins, as is often the case with education and occupation, estimating rank-based mobility measures is challenging for two reasons.[1] First, when bins are large, the relationship between parent and child outcomes *within* bins may be important for mobility but is not observed by the researcher. In India, for example, 60% of fathers of the 1950s birth cohort had zero years of education. Any latent differences in opportunity within the bottom half of the distribution are thus not observed. Conventional mobility estimates in this context implicitly make the strong assumption that a child growing up in the 10th percentile family has the same set of opportunities as a child growing up in the 50th percentile family. Second, it is difficult to compare mobility estimates across contexts with different bin boundaries. For example, if education is rising over time, then uneducated parents represent lower ranks with each successive cohort. Constructing comparable quantile bins requires either strong assumptions or a set identification approach where parameters of interest may at best be bounded.[2]

In this paper, we develop a method that generates unbiased bounds on measures of rank mobility when ranks are coarsely observed. We show that when we weaken implicit assumptions about within-bin parent-child outcome correlations, we can at best identify a set of parameters containing the true mobility estimate. In some cases the bounds are tight, but in others they are too wide to be informative. This leads us to propose a new measure of

[1] Recent studies of intergenerational mobility focusing on education and occupation include Black et al. (2005), Long and Ferrie (2013), Güell et al. (2013), and Wantchekon et al. (2015).

[2] Set identification, which uses structural assumptions to refine worst-case bounds on problems where point identification is impossible, is summarized in Manski (2003), Tamer (2010), and Ho and Rosen (2015).

intergenerational mobility that can be tightly bounded even when data are extremely coarse.

We begin with the assumption that the rank bin data represent a latent, continuous rank distribution that is observed only in coarse intervals.[3] We can then develop a method to analytically bound the conditional expectation function of child rank given latent parent rank (the CEF henceforth).[4,5] The bounds can be further tightened with assumptions about the curvature of the CEF, though solving this problem requires numerical optimization. In the limit, restricting the curvature to zero is equivalent to assuming that the canonical rank-rank gradient fully describes the parent-child distribution. The method allows us to bound the expected child outcome given parent rank at any point in the rank distribution, even if that rank is observed only in a coarse bin.[6] To validate the method, we simulate interval censoring in Danish income data where the full rank distribution is known and we show that we can recover sensible bounds given just the censored data.

From the bounds on the child CEF, we can bound any mobility statistic derived from the CEF, including the rank-rank gradient and newer measures like absolute upward mobility (Chetty et al., 2014a). When the rank bins are large, we show that both of these measures may have bounds that are too wide to be informative.

We propose a new measure of mobility that can often be bounded tightly even with severely interval censored data. We define *interval mobility*, or $\mu_a^b$, as the expected outcome of children born to parents who occupy positions between $a$ and $b$ in the parent rank distribution. A useful instance of this measure is $\mu_0^{50}$, which describes the expected outcome of a child born

[3]Parent education data in mobility studies most often reports not even years of education, but the highest major level of schooling completed, such as primary or middle. The latent educational rank could represent the continuous time spent in school (i.e. the number of days with a teacher present), as well as parent inputs into human capital accumulation.

[4]Specifically, we prove sharp analytical bounds on such a CEF $E(Y|x)$ when $x$ is interval censored and has a known distribution. These bounds substantially improve upon those established by Manski and Tamer (2002) for $x$ variables with unknown distributions. When $x$ is uniformly distributed, as when it represents parent ranks, the bounds are particularly parsimonious. Chetty et al. (2017a) numerically calculate bounds on the probability of children having higher income than their parents, in a context where they observe only unmatched incomes from two generations.

[5]To account for sampling variation, we generate bootstrap confidence sets for all estimates (Imbens and Manski, 2004; Tamer, 2010).

[6]We focus on addressing interval censoring in parent rank data, which is typically a greater source of bias than censoring in child rank data, but we address the latter in Section IV.F and Appendix C.

in the bottom half of the distribution; we call this *upward interval mobility.* This measure is similar in spirit to absolute upward mobility, but we show that upward interval mobility can be tightly bounded under severe interval censoring, while absolute upward mobility cannot. We prove sharp analytical bounds on $\mu_a^b$, and show how to obtain numerical bounds on $\mu_a^b$ given additional curvature restrictions.

We apply these methods to the calculation of educational mobility in India, using new administrative data on the education of all fathers and sons in the country.[7] Education ranks are very coarsely observed: 60% of parents of the oldest generation (and 38% of parents of the youngest) report a level of education in the lowest category. Comparable numbers are common in studies of educational and occupational mobility (see Table 1).

Using our new measure of upward interval mobility, we show that the expected ranks of children in the bottom half of the education distribution have barely changed since the 1950s. In contrast, the bounds on both the traditional rank-rank gradient and on absolute upward mobility are only moderately informative and are consistent with both large gains and small losses in mobility since the 1950s. If the coarse data problem were ignored, the naive approach to estimating the rank-rank gradient would suggest small but unambiguous mobility gains over the same period.[8]

We then separately examine upward mobility for Scheduled Castes and Scheduled Tribes (SC/STs) and for Muslims. These groups each represent about 15% of the population and have been historically marginalized, though SC/ST disadvantage has been targeted by many programs and policies since independence (Munshi and Rosenzweig, 2006; Ito, 2009; Hnatkovska et al., 2012). We focus on upward interval mobility, as it is the only measure that we can bound narrowly. There are three main findings. First, once we separate SC/STs and Muslims, upward mobility for the remainder of the population is comparable to the

[7]While mobility with respect to daughters and mothers is equally relevant, as is often the case, we are constrained because the only matched parent-child pairs that we observe in our data are fathers and sons.

[8]Because the fully supported rank distribution is uniform, the rank correlation coefficient is equivalent to the coefficient from a regression of child rank on latent parent rank. This equivalence does not necessarily hold with interval rank data.

United States, and has not changed over the last thirty years. Second, upward mobility has improved significantly for SC/STs since the 1960s; the expected educational rank of an SC/ST child born into the bottom half of the distribution has risen from 33 in the 1960s to 36 in the 1980s. Third, this change is almost exactly mirrored by a decline in upward mobility among Muslims, which has fallen from 31.5 to 29 over the same period. Results are robust to different data construction methods, and are not significantly affected by survivorship bias.

These findings imply that virtually all of the mobility gains in India over recent decades have accrued to SC/STs, and none have come at the expense of the upper caste population, in spite of significant political mobilization against affirmative action by higher caste groups in recent years. For non-SC/STs, there is no evidence that economic liberalization has substantially increased opportunities for those in the lower half of the rank distribution to attain higher relative social position, and for Muslims these opportunities have substantially deteriorated.[9]

These patterns have to our knowledge not been identified because earlier studies have either (i) focused on absolute outcomes (such as consumption), which are rising for all groups due to India's substantial economic growth (Maitra and Sharma, 2009; Hnatkovska et al., 2013); or (ii) compared subgroups using the parent-child outcome correlation or regression coefficient, which describe the outcomes of subgroup members relative to their own group, rather than to the national population (Hnatkovska et al., 2013; Emran and Shilpi, 2015; Azam and Bhatt, 2015). Studies on affirmative action in India have identified improvements in SC/ST access to higher education but have not examined impacts on Muslims (Frisancho Robles and Krishna, 2012; Bagde et al., 2016); our findings point to the importance of studying the effects of such policies on other marginalized groups.[10]

In summary, this paper makes four main contributions. First, we prove new analytical

[9]Absolute outcomes for all social groups have been improving systematically due to India's substantial economic growth over the study period. We focus on the ability of individuals to rise to a higher *relative* position in society than their parents exactly because it holds constant the level shift in outcomes due to growth. However, absolute measures of mobility are also of interest and can be bounded using our method.

[10]In a similar finding, Bertrand et al. (2010) find that when Indian colleges intentionally select lower caste students, they admit fewer women.

bounds on a conditional expectation function $E(y|x)$ when $x$ is interval censored with a known distribution, and we show how to compute numerical bounds with additional curvature restrictions. Second, we develop a method to compare expected child outcomes at constant parent rank intervals even when parent ranks are not observed at exactly those intervals. This makes it possible to compare the educational mobility of subgroups with respect to the entire population, which was previously difficult. Third, we propose a new measure of upward mobility that can be tightly bounded even when parent ranks are measured very coarsely. Fourth, we present new national and subgroup estimates of upward mobility in India that are robust to coarse parent data; to our knowledge, the declining intergenerational mobility of Muslims has not been previously established.[11,12]

The problem of coarse data is common, especially in developing countries where mobility studies frequently use education as an indicator of status and a large share of older generations has not completed primary school. Our method is also applicable in wealthier countries where a large share of the population may be in a topcoded education bin.[13] Larger or more representative samples do not mitigate the challenge of coarse measurement of rank, and additional covariate data on parents of older generations are often unavailable.

Our approach to interval rank data can also be helpful in other contexts where researchers need to construct comparable rank bins across time or space. This problem arises, for instance, in the study of disparities in health outcomes across the education distribution. Case and Deaton (2015, 2017) estimate changes in mortality over time among groups with a high school degree or less; they do not look specifically at less educated groups because their population share, and thus their relative social rank, changes over time. Bound et al. (2015)

[11]The persistent low absolute levels of economic outcomes for Muslims have been noted by researchers and in occasional government reports but have not been a focus of government policy (Sachar Committee Report, 2006; The Economist, Oct 29, 2016).

[12]Our mobility estimates are not intended to be interpreted as causal effects of parent schooling on child schooling. They are descriptive measures of the degree of equality of opportunity in a society.

[13]In one mobility study from Sweden, for example, 40% of adoptive parents were topcoded with 15 or more years of education (Björklund et al., 2006). Studies on the persistence of occupation across generations also frequently use a small number of categories and face a similar challenge when the occupational structure changes significantly over time, as it has with farm work in the United States. See, for example, Long and Ferrie (2013), Xie and Killewald (2013) and Guest et al. (1989).

show that failure to control for changing sizes of rank bins leads to biased mortality estimates. Our method can bound mortality measures on arbitrary education rank intervals that are not directly observed.

Our paper proceeds as follows. Section II describes current approaches to estimating intergenerational mobility, with a focus on estimates using coarse data. Section III describes the Indian data. Section IV describes our method for calculating nonparametric bounds on the child CEF and presents the new measure of upward interval mobility. We validate the methods using data from Denmark, and work through an example with data from a single decadal cohort of Indian children. Section V applies our method to the estimation of changes in mobility over time and estimating differences in mobility between social groups in India. Section VI concludes. Stata and Matlab code to calculate all measures used in this paper are available on the corresponding author's web site.

## II Background: Theory and Empirics of Mobility Measurement

When intergenerational mobility is low, the social status of individuals is highly dependent on the social status of their parents (Solon, 1999). In more mobile societies, individuals are less constrained by their circumstances at birth. There is a growing literature on the variation in intergenerational mobility across countries, across groups within countries, and across time.[14]

The first generation of intergenerational mobility studies described matched parent-child outcome distributions with a single linear parameter, typically using either: (i) the correlation coefficient between children's earnings and parents' earnings; or (ii) the parent-child earnings elasticity, *i.e.* the coefficient from a regression of children's log earnings on parents' log earnings (Solon, 1999; Black and Devereux, 2011).[15] These measures are easy to calculate and interpret, and they form the basis of studies in dozens of countries. However, the gradient estimator is not well-suited for between-group comparisons, because the subgroup

[14]See Hertz (2008) for cross-country comparisons, and Solon (1999), Corak (2013), Black and Devereux (2011), and Roemer (2016) for review papers.

[15]These measures are related through the variance of parents' and children's outcomes: in the absence of additional regressors, $\beta = \rho \frac{\sigma_{children}}{\sigma_{parents}}$, where $\sigma$ denotes the standard deviation of lifetime earnings, $\rho$ is the correlation coefficient and $\beta$ is the parent-child regression coefficient.

gradient measures children's outcomes against better off members of their own group; a subgroup can therefore have a lower gradient (suggesting more mobility) in spite of having worse outcomes at every point in the parent distribution.

Recent studies have instead analyzed the parent-child income and education *rank* distributions (Boserup et al., 2014; Chetty et al., 2014a; Hilger, 2016; Bratberg et al., 2015), with both parent and child ranks calculated within the children's cohorts. Since the parent and child rank distributions are uniform, the parent-child rank correlation coefficient and rank regression coefficient (the gradient) are identical, and the mobility statistic is invariant to changes in the variance of the parent-child outcome distribution.[16] Parent-child rank distributions are also more easily compared across countries than level distributions.

Studies drawing on rich administrative data such as tax records have examined the entire nonparametric conditional expectation function of child rank given parent rank (the CEF). Studies have highlighted that the CEF is often non-linear (the United States being an important exception): both mobility and changes in mobility may vary across the parent rank distribution.[17] Many useful statistics of mobility can be generated from the parent-child rank CEF. First, the rank-rank gradient is the slope of the best linear approximation of the CEF. Second, Chetty et al. (2014a) propose a measure that describes the value of the CEF at any given rank, which they call absolute mobility. Absolute mobility at the $i$-th percentile, which we denote by $p_i$, is the expectation of a child's rank if the child's parent is at the $i$-th percentile.[18] Chetty et al. (2014a) pay particular attention to $p_{25}$, which describes the

[16]Importantly, this invariance does not hold if the rank data are binned differently for parents and children, or across different estimations.

[17]The CEF of child income rank given parent income rank is largely linear in the United States (Chetty et al., 2014b), but has important nonlinearities in Denmark (Boserup et al., 2014), Norway and Sweden (Bratberg et al., 2015), especially in the tails. The CEF of child college quality given parent income rank in the United States is significantly convex (Chetty et al., 2014a), as are the CEFs of log income in Denmark, Finland and Norway (Bratsberg et al., 2007). We are not aware of any studies from developing countries with sufficient data to report a fully supported rank CEF, though we show below that the education rank CEF is India is nonlinear.

[18]Chetty et al. (2014a) use the term "absolute mobility" because this measure does not depend on the value of the CEF at any other point in the parent rank distribution, distinguishing it from the rank-rank correlation which they describe as a *relative* mobility measure. Other authors use the term "absolute mobility" to describe the set of mobility measures which use child levels as outcomes rather than child ranks. Because this study is focused on child rank outcomes, we avoid the term "absolute mobility" except

expected rank of the median child born in the bottom half of the distribution.

## II.A Educational Mobility and Income Mobility

Education data have been widely used for mobility studies for three reasons: (i) matched parent-child education data are more widely available than matched parent-child income data; (ii) education may be measured more precisely than income among the very poor; and (iii) they are less likely to be influenced by life-cycle bias.[19]

Representative matched parent-child income data are frequently unavailable to the researcher. In developing countries, there are few administrative income records, and if they do exist, they rarely if ever date back long enough to measure parent and child incomes at similar ages.

Even when matched parent and child income data are available (for example, for the set of parents and children who are coresident), measurement error problems are significant. Transitory incomes are noisy estimates of lifetime income, subsistence consumption is difficult to measure, and many individuals report zero income; these problems are exacerbated among the rural poor.[20] As a result, studies of intergenerational mobility in developing countries often proxy lifetime income and opportunity with the level of education. This approach has been validated in countries where both approaches are possible; intergenerational educational mobility is highly correlated with intergenerational income mobility (Solon, 1999). In developing countries in particular, educational attainment may be a better proxy for lifetime income than a single observation of transitory income.

Focusing on education also avoids the problem of life-cycle bias in income measurement. The income or income rank of a young person may not be a good proxy for her lifetime

when making reference to the specific measure used by Chetty et al. (2014a).

[19] A sample of studies examining educational mobility includes Solon (1999), Black et al. (2005) and (Restuccia and Urrutia, 2004). More are summarized in Black and Devereux (2011)

[20] Intergenerational income mobility studies either focus on individual wages or total household earnings, both of which are problematic. Given that transitioning out of agricultural work and into wage work is a central predictor of consumption, restricting the sample to wage earners has obvious deficiencies. But for the significant set of households where parents and children are coresident, total household earnings include earnings of the parent, which biases downward the measurement of mobility.

income, because high permanent income individuals may spend more time in school and thus have lower income than their peers when young. In contrast, the level of education rarely changes later in life.

Estimates of educational mobility do have one important drawback, which is the focus of this paper: outcomes are typically reported in a small number of categories. Even when years of education are specifically measured, they are bunched at school completion levels. A large share of individuals report zero completed years of education — in fact, this number exceeds half of the population in many poor countries. The small number of categories makes it difficult to compare the outcomes of similarly ranked individuals across time periods with changing rank boundaries. For example, consider the top rows of the father-son transition matrices presented in Appendix Table A1. Among Indian sons born in the 1950s, 60% have fathers with zero education. In the 1980s, the data provide no comparably ranked parent group; 38% of fathers have zero education, 49% have zero or less than primary, and 67% have primary or less. Researchers often compare transition matrices across countries by focusing on quartile or quintile transition matrices; but we are aware of no established methodology for generating such matrices rank data with arbitrary rank bins. Researchers in this circumstance often implicitly assume parent and child outcomes are independent *within* bins of the transition matrix. In the 1950s Indian son cohort, this amounts to assuming that an Indian son born at the 10th percentile of the opportunity distribution has equivalent opportunity to a son born at the 50th percentile, which is unlikely to be the case.

To make it possible to compare similarly sized rank bins in the parent-child education distribution, we assume that every parent and child can be characterized by a latent continuous education rank, which is observed in censored intervals by the researcher. The unobserved component of the education rank could include partially completed years of schooling or the number of days spent in school with a teacher present. In this framework, we show that it is possible to place bounds on the latent parent-child rank CEF given the observed data, and that we can generate informative mobility statistics describing mobility even in the parts of the distribution where bins are the most coarse. This framework allows us to consider, for example, the possibility that a child born at the 10th percentile of the education distribution has a different expected outcome from a child born at the 50th percentile, even if we do not observe these ranks directly.[21]

Table 1 describes a set of recent studies of intergenerational mobility from several rich and poor countries.[22] For each study, we report the range of birth cohorts of sample children, the number of education bins used (where reported), and the share of the parents reporting zero education. Several of the studies observe education in fewer than ten bins, the population share in the bottom bin is often above 20%, and sometimes it is above 50%.[23] Even in the youngest cohorts, a large share of parents have not attained primary school. By contrast, in developed countries, the size of the *top* bin may be too large to develop tight bounds on the CEF. For example, Björklund et al. (2006) study the education of adopted children and report that the size of the top education category (15 years or more) for adoptive parents is 40%.

Long run studies of mobility in developed countries often rely upon occupation as a measure of social status (Guest et al., 1989; Long and Ferrie, 2013). When a large share of the population is classified in a given occupation (for instance, unskilled labor), any intergenerational persistence of occupation at an unobserved and more finely detailed level will bias upward mobility estimates based on broad categories. Significant differences in the population share in each occupation across settings also complicates mobility comparisons (Xie and

[21]The interval censoring assumption is not totally innocuous, because observed years of education describe latent continuous education rank with some error. For example, some middle-school educated people could in fact have less education than some primary-school educated people, because of variation in school quality. Measurement error like this will bias mobility estimates upward. While we do not address the measurement error problem in this paper, we note that studies using years of education will equally overestimate mobility if years of education do not accurately reflect opportunity. We also focus on comparison of mobility across time and across subgroups in India, rather than across countries where differences in measurement error may be more serious.

[22]We specifically selected a set of studies where coarse data is likely to be an important factor. This list notably excludes several recent studies using tax data from developed countries, where the parent rank distribution is observed without substantial interval censoring.

[23]The studies that we found from Africa did not report the size of the bottom bin, but estimates of primary completion suggest that it is large. In 1980, the interquartile range of the primary enrollment rate was (0.48, 0.98) and the interquartile range of the youth literacy rate was (0.27, 0.64) (World Development Indicators). These are indicators of children's education, so the number of parents without primary education is considerably higher.

Killewald, 2013). The method in this paper can be applied to intergenerational occupational mobility only to the extent that occupational categories can be monotonically ranked, which may be possible in some categorizations but not in others. Finally, studies of intergenerational income mobility often rely on interval censored data; the literature on the British Cohort Study (BCS) is a well-known example. Our method can be directly applied to generate estimates that take the interval censoring into account.[24] Other studies using coarse data to estimate long-run mobility include Lindahl et al. (2012), Clark (2012), and Clark (2014).

### II.B Context: Educational Mobility in India

While intergenerational mobility is of interest around the world, India's caste system and high levels of inequality make it a particularly important setting for such work. India's caste system is characterized by a set of rules that expressly inhibit intergenerational mobility by preventing individuals from taking up work outside of their caste's traditional occupation and from marrying outside of their caste. While some have argued that economic growth is making old social and economic divisions less important to the economic opportunities of the young, caste remains an important predictor of economic opportunity.[25] India's Muslims, who are approximately equal in population to the Scheduled Castes and Scheduled Tribes (SC/STs), have also been historically marginalized, albeit through different mechanisms. Since independence, several major political parties have successfully rallied around Hindu nationalist policies which explicitly discriminate against Muslims and Muslim religious practices. The ruling federal coalition at the time of writing (the BJP) arose out of the Hindu nationalist movement. Muslims have also frequently been targeted by mob violence. The last 30 years have seen tremendous growth in market opportunities in India as well as in the availability and level of education, but many have argued that this growth has been unequal

[24] We generated intergenerational income mobility estimates using the BCS, and found that bounds on elasticities tightly contain the conventionally estimated parameters, because the income bins are sufficiently narrow in most of the distribution. Bounds on absolute mobility at the 25th percentile are coarser. Results are available from the authors.

[25] For examples, see Munshi and Rosenzweig (2006), Ito (2009), Hnatkovska et al. (2013), and Mohammed (2016).

and is leaving many behind (Dreze and Sen, 2013; Field et al., 2016). Whether economic progress can overcome traditional hierarchies of social class and religion is a central question for both India and the broader world.

Increasing intergenerational mobility in India has also been a major objective of the Indian government since independence, leading to reservations of educational and political positions for members of India's disadvantaged groups, among other policies. Group-targeted policies have nearly all focused on SC/STs, and to our knowledge no major policies have had the goal of specifically ameliorating Muslim disadvantage. Understanding how mobility has changed for these population groups is thus an essential component of understanding secular trends in intergenerational mobility in India.

## III Data

To estimate intergenerational educational mobility in India, we draw on two databases that report matched parent-child educational attainment. The first is an administrative census dataset describing the education level of all parents and their coresident children. Because coresidence-based intergenerational mobility estimates may be biased, we supplement this with a second dataset with a representative sample of non-coresident father-son pairs.[26] We focus on fathers and sons because we do not have data on non-coresident mothers and/or daughters. This section describes the two datasets.

The Socioeconomic and Caste Census (SECC) was conducted in 2012, to collect demographic and socioeconomic information determining eligibility for various government programs.[27] The data was posted on the internet by the government, with each village and urban neighborhood represented by hundreds of pages in PDF format. Over a period of two years,

[26]Emran et al. (2017) argue that the coresidence bias for intergenerational mobility depends on how mobility is measured, and that the father-son correlation coefficient has only a small bias. In our own estimates using a similarly powered test, we found a similarly small point estimate to Emran et al. (2017), but the 95% confidence interval for the estimate leaves open the possibility of substantial bias in coresidence-based estimates. Further, the correlation measure is less useful for evaluating differences across groups. It is thus important to work with estimates that include non-coresident parent-child pairs.

[27]It is often referred to as the 2011 SECC, as the initial plan was for the survey to be conducted between June and December 2011. However, various delays meant that the majority of surveying was conducted in 2012. We therefore use 2012 as the relevant year for the SECC.

we scraped over two million files, parsed the embedded data into text, and translated the text from twelve different Indian languages into English. The individual-level data that we use describe age, gender, an indicator for Scheduled Tribe or Scheduled Caste status, and relationship with the household head. Assets and income are reported at the household rather than the individual level, and thus cannot be used to estimate mobility.[28] The SECC provides the education level of every parent and child residing in the same household. Sons who can be matched to fathers through coresidence represent about 85% of 20-year-olds and 7% of 50-year-olds. Education is reported in seven categories.[29] To ease the computational burden of the analysis, we work with a 1% sample of the SECC, stratified across India's 640 districts.

We supplement the SECC with data from the 2011-2012 round of the India Human Development Survey (IHDS). The IHDS is a nationally representative survey of 41,554 households in 1,503 villages and 971 urban neighborhoods across India. Crucially, the IHDS solicits information on the education of fathers of household heads, even if the fathers are not resident, allowing us to fill the gaps in the SECC data. Since the SECC contains data on all coresident fathers and sons, our main mobility estimates use the IHDS strictly for non-coresident fathers and sons. IHDS contains household weights to make the data nationally representative; we assign constant weights to SECC, given our use of a 1% sample. By appending the two datasets, we can obtain an unbiased and nationally representative estimate of the joint parent-child education distribution.[30] IHDS reports neither the education of non-coresident mothers nor of women's fathers, which is why our estimates are restricted to fathers and sons.

IHDS records completed years of education. To make the two data sources consistent, we recode the SECC into years of education, based on prevailing schooling boundaries, and we downcode the IHDS so that it reflects the highest level of schooling completed, *i.e.*, if someone reports thirteen years of schooling in the IHDS, we recode this as twelve years,

[28] Additional details of the SECC and the scraping process are described in Asher and Novosad (2017).

[29] The categories are (i) illiterate; (ii) literate without primary (iii) primary; (iv) middle; (v) secondary (vi) higher secondary; and (vii) post-secondary.

[30] We verified that IHDS and SECC produce similar point estimates for the coresident father-son pairs that are observed in both datasets. Point estimates from the IHDS alone (including coresident and non-coresident pairs) match our point estimates, albeit with larger standard errors.

which is the level of senior secondary completion.[31] The loss in precision by downcoding the IHDS is minimal, because most students exit school at the end of a completed schooling level.

The public version of the SECC does not report religion, so our subgroup estimates are generated entirely from the 2011 IHDS, with both coresident and non-coresident parent-child pairs included. Because religion is a separate classification from caste, some Muslims are also members of Scheduled Castes and Tribes; some of the latter have converted to Islam. IHDS reports caste and religion together, and SC/ST Muslims are reported only as SC/STs. However, fewer than 2% of Muslims are SC/STs (Sachar Committee Report, 2006), so this distinction is unlikely to affect our estimates.

We estimate changes in mobility over time by examining the joint distribution of fathers' and sons' educational attainment for sons in different birth cohorts. All outcomes are measured in 2012, but because education levels only rarely change in adulthood, these measures capture educational investments made decades earlier. We use decadal cohorts reflecting individuals' ages at the time of surveying. To allay concerns that differential mortality across more or less educated fathers and sons might bias our estimates, we replicate our analysis on the *same* birth cohorts using the IHDS 2005. By estimating mobility on the same cohort at two separate time periods, we identify a small survivorship bias for the 1950-59 birth cohort (reflecting attrition of high mobility dynasties), but zero bias for the cohorts from the 1960s forward. Our results of interest largely describe trends from the 1960s forward (in part because standard errors are largest for the 1950s cohort, making inference more difficult), so survivorship bias among the oldest cohorts does not influence any of our conclusions.

The oldest cohort of sons that we follow was born in the 1950s and would have finished high school well before the beginning of the liberalization era. The cohorts born in the 1980s would have completed much of their schooling during the liberalization era. The youngest cohort in this study was born in 1989; cohorts born in the 1990s may not have completed

[31] We code the SECC category "literate without primary" as two years of education, as this is the number of years that corresponds most closely to this category in the IHDS data, where we observe both literacy and years of education. Results are not substantively affected by this choice.

their education at the time that they were surveyed and were excluded. Appendix Figure A1 shows the educational attainment of fathers and sons by cohort; all social groups show dramatic gains in educational attainment over the last forty years. SC/STs have closed about a third of the gap with the non-Muslim, non-SC/ST population, while the gap between Muslims and this population has closed only marginally.

## IV Estimating Bounds on Intergenerational Mobility

This section describes the main contribution of the paper. Here, we describe a method to calculate bounds on a range of intergenerational mobility statistics that take into account interval censoring in the parent rank distribution. We first show that interval rank data can be used to bound the conditional expectation of child rank given parent rank. The bounds are sharp and depend only on the assumption of a weakly monotonic CEF; imposing an additional restriction on the CEF curvature can yield tighter bounds but requires numerical optimization to solve. Given CEF bounds, we can then bound many measures of mobility. We focus first on the rank-rank gradient and absolute upward mobility (Chetty et al., 2014a), and then show that a new measure, upward interval mobility, generates particularly tight bounds under interval data. We focus on the CEF of child rank given parent rank, but the method can be equally applied to any CEF of interest. For example, we can directly calculate bounds on the probability that a child from a given part of the parent rank distribution attains the top quintile by substituting $E(Y|x)$ with $E(P(Y > 80)|x)$.

We walk through the method with sample data from India. We then validate it by simulating interval censoring in uncensored data from Denmark and recovering bounds on the original CEF. Finally, we discuss the extension of the method to take into account interval data on children as well as parents.

### IV.A Nonparametric Inference on Mobility with Interval Data

Define the child rank as $y$ and the parent rank as $i$; the conditional expectation function of child rank given parent rank is $Y(i) = E(y|i)$. Let the function $Y(i)$ be defined on $i \in [0, 100]$, and assume $Y(i)$ is integrable.[32]

With interval data, we do not observe $i$ directly, but only that it lies in one of $K$ bins.[33] Define the expected child outcome in the $k^{th}$ bin as

$$r_k = E\left(y|i \in [i_k, i_{k+1}]\right) = \frac{1}{i_{k+1} - i_k} \int_{i_k}^{i_{k+1}} Y(i)di,$$

where $i_k$ and $i_{k+1}$ define the bin boundaries. For the outer rank bins, $i_0 = 0$ and $i_{K+1} = 100$. Further define the expected child outcomes in the intervals directly above and below the intervals of interest as $r_{k+1} = E\left(y|i \in [i_{k+1}, i_{k+2}]\right)$ and $r_{k-1} = E\left(y|i \in [i_{k-1}, i_k]\right)$, if they exist. Define $r_0 = 0$ and $r_{K+1} = 100$. We observe the sample analogs to these expressions in the data, and denote the mean outcome in bin $k$ with $\overline{r}_k$.[34]

Panel A of Figure 1 depicts the setup, using data from Indian men born in the 1960s. Parent education is observed in seven bins, representing the highest level of schooling attained by each father, with the vertical lines showing the bin boundaries. The bottom bin comprises 57% of fathers, all of whom report zero education. The points in the graph show the mean child rank conditional on having a parent in a given bin, which is $\overline{r}_k$.[35] The two functions in Panel A show two (of many) possible nonparametric CEFs, each of which fit the sample data equally well. Note that both of these functions have the same mean in the first bin, even though neither function passes through that mean at the bin midpoint. In contrast, the conventional linear approach to mobility estimation implicitly assumes that the true CEF passes through this mean.[36] Figure 1B shows the rank-rank gradients that would be fitted to each of these CEFs (dashed lines), as well as the conventional rank-rank gradient that would be estimated from the bin midpoints. If the rank-rank gradient is understood as

[32] The results in this subsection apply to any monotonic function $Y(i)$ but we use the language of parent and child rank for concreteness.

[33] For notational simplicity, we assume that all parent ranks are censored using the same set of boundaries.

[34] In this paper we focus only on child ranks as outcomes but in principle any outcome of interest can be used.

[35] We calculate the mean child rank using the midpoint of each interval in which child rank is observed. We loosen this assumption in Section IV.F.

[36] A naive polynomial fit to the points in the graph would not even fit the underlying data because of Jensen's Inequality.

a linear approximation to a potentially nonlinear CEF, then all of these gradients (among others) can fit the underlying data equally well.[37] If the true CEF is nonlinear within bins, then conventional gradient estimates may be biased due to Jensen's inequality. We aim to identify the set of all CEFs that satisfy the data under some set of assumptions.

Manski and Tamer (2002) derive bounds on $E(y|i)$ given interval measurement of $i$. The essential structural assumption that constrains the child outcome is Monotonicity (M):

$$E(y|i) \text{ must be weakly increasing in } i. \tag{Assumption M}$$

In our context, this assumption states that a child born to a higher ranked parent will on average attain an equal or higher rank in the outcome distribution.[38] Manski and Tamer (2002) also introduce the following Interval (I) and Mean Independence (MI) assumptions. For $i$ which appears in the data as lying in bin $k$,

$$P(i \in [i_k, i_{k+1}]) = 1. \tag{Assumption I}$$

$$E(y|i \text{ is interval censored}) = E(y|i). \tag{Assumption MI}$$

Assumption $I$ says that the rank of all people who report education ranks in category $k$ are actually in bin $k$. Assumption $MI$ states that censored observations are not different from non-censored observations; this always holds in our context because all of the data are interval censored.

Suppose these IMMI assumptions hold. Manski and Tamer's Proposition 1 provides the following sharp bounds on $E\left(y|i\right)$, given our setup with fully interval censored data:

$$r_{k-1} \leq E(y|i) \leq r_{k+1} \tag{Manski-Tamer bounds}$$

[37]The rank-rank gradient remains useful as a single summary statistic for the entire distribution even if the underlying data are nonlinear, as is clear in this case.

[38]Even a system that obtains perfect equality of opportunity will have this property, and it is observed in every empirical observation of the parent-child income and education distributions that we are aware of.

We can improve upon these bounds because we have information about the distribution of $i$. In our setting, parent ranks are uniformly distributed. Suppose further that the following Uniformity assumption (U) holds:

$$i \sim U(0, 100) \qquad \text{(Assumption U)}$$

where $U$ is the uniform distribution.

Because $i$ is uniformly distributed, we know that:

$$E\left(i | i \in [i_k, i_{k+1}]\right) = \frac{1}{2}\left(i_{k+1} - i_k\right). \qquad \text{(IV.1)}$$

We derive the following proposition.

**Proposition 1.** *Let $i$ be in bin $k$. Under assumptions IMMI (Manski and Tamer, 2002) and U, the following bounds on $E(y|i)$ are sharp:*

$$\begin{cases} r_{k-1} \leq E(y|i) \leq \frac{1}{i_{k+1}-i}\left(\left(i_{k+1} - i_k\right) r_k - \left(i - i_k\right) r_{k-1}\right), & i < k^* \\ \frac{1}{i-i_k}\left(\left(i_{k+1} - i_k\right) r_k - \left(i_{k+1} - i\right) r_{k+1}\right) \leq E(y|i) \leq r_{k+1}, & i \geq k^* \end{cases}$$

*where*

$$k^* = \frac{i_{k+1} r_{k+1} - \left(i_{k+1} - i_k\right) r_k - i_k r_{k-1}}{r_{k+1} - r_{k-1}}.$$

At the interval endpoints ($i = i_k$ or $i = i_{k+1}$), these bounds reduce to the bounds from Manski and Tamer (2002).

The proposition is obtained from the insight that the expected rank at a point $I$ in bin $k$ will only be minimized if all points in bin $k$ to the left of $I$ have the same value. Since all points to the right of $I$ are constrained by the outcome value in the subsequent bin $k+1$, $E(y|i = I)$ will need to rise above the Manski and Tamer (2002) lower bound as $I$ increases in order to match the bin mean. Appendix B provides additional intuition, and a proof of this proposition not only for the uniform case, but for any case where the distribution of $i$ is known but not necessarily uniform. In the mobility literature, the more general result could be of interest when dealing with censored income data, given an assumption that the top tail distribution is Pareto, for instance.

Using the Indian data, Figure 2 compares Manski and Tamer (2002) bounds to those obtained under the additional assumption of uniformity. The new bounds are a significant improvement, especially in places where the data are particularly coarse.

Under the same set of assumptions, we can additionally obtain bounds on several mobility statistics that are simple functions of the CEF. We clearly bound absolute mobility at the $i^{th}$ percentile, defined by Chetty et al. (2014a) as the expected outcome of a child born to parents at percentile $i$, denoted $p_i$. This is exactly $E(y|i = I)$ and is bounded by Proposition 1. Chetty et al. (2014a) focus on $p_{25}$, which is the expected outcome of the median child born to parents in the bottom half of the distribution.

We propose a new measure of mobility, which is the expected outcome of a child born to a parent with rank between $a$ and $b$, or $E(y|i \in [a, b])$; we call this measure *interval mobility* and denote it $\mu_a^b$. An analogous measure to $p_{25}$ is $\mu_0^{50}$, which describes the expected outcome of the *mean* child in the bottom half of the distribution; we call this measure *upward interval mobility*. Put otherwise, define $\mu_a^b$ as

$$\mu_a^b = \frac{1}{b-a} \int_a^b E(y|i)di. \tag{IV.2}$$

Let $Y_i^{max}$ be the analytical upper bound on $E(y|i)$, given by Proposition 1. Let $Y_i^{min}$ be the analytical lower bound on $E(y|i)$. We obtain the following bounds on $\mu_a^b$.

**Proposition 2.** *Let $b \in [i_k, i_{k+1}]$ and $a \in [i_h, i_{h+1}]$. Let assumptions IMMI (Manski and Tamer, 2002) and U hold. Then the following bounds are sharp:*

$$\begin{cases} Y_b^{min} \leq \mu_a^b \leq Y_a^{max} & h = k \\ \frac{r_h(i_k - a) + Y_b^{min}(b - i_k)}{b-a} \leq \mu_a^b \leq \frac{Y_a^{max}(i_k - a) + r_k(b - i_k)}{b-a} & h+1 = k \\ \frac{r_h(i_{h+1} - a) + \sum_{\lambda=h+1}^{k-1} r_\lambda(i_{\lambda+1} - i_\lambda) + Y_b^{min}(b - i_k)}{b-a} \leq \mu_a^b \leq \frac{Y_a^{max}(i_{h+1} - a) + \sum_{\lambda=h+1}^{k-1} r_\lambda(i_{\lambda+1} - i_\lambda) + r_k(b - i_k)}{b-a} & h+1 < k. \end{cases}$$

The proof is in Appendix B.

Interval mobility measures are often more tightly bounded than absolute mobility for a given region of the parent rank distribution, especially when the bottom rank bin is large. Appendix Figure A2 makes this point, showing absolute mobility and interval mobility for progressively lower ranges in the parent rank distribution, with each set of estimates centered on the same parent bin. The bounds on interval mobility are consistently tighter than those on absolute mobility, though as we move to the bottom of the distribution, the advantage is minimal.

Similarly, upward interval mobility ($\mu_0^{50}$) generates much tighter bounds than absolute upward mobility ($p_{25}$) when the bottom rank bin is large. Using the Indian data above, we find $\mu_0^{50} \in [34.8, 38.7]$, and $p_{25} \in [22.2, 54.5]$. For intuition, observe that if $\delta \in [a, b]$, $\mu_a^b$ can be decomposed into $\frac{\delta-a}{b-a}\mu_a^\delta + \frac{b-\delta}{b-a}\mu_\delta^b$. For intervals exactly corresponding to bin boundaries, $\mu_k^{k+1}$ is determined and must satisfy $\mu_k^{k+1} = r_k$. Thus if $\mu_0^{50}$ extends over several intervals where the mean is known, $\mu_0^{50}$ has less room to vary.

Bounds on other mobility statistics may be difficult to calculate analytically, but can be defined as the set of solutions to a pair of minimization and maximization problems that take the following structure. Let us write the conditional expectation function in the form $Y(i) = s(i, \gamma)$, where $\gamma$ is a finite-dimensional vector that lies in parameter space $G$ and serves to parameterize the CEF through the function $s$. This parameterization is not restrictive: in our numerical optimization below, we will define $\gamma$ as a vector of 100 discrete values of the CEF in each single rank interval. However, it could also take a more parsimonious parametric form. Define $\Gamma$ as the set of feasible parameterizations of the CEF that obey monotonicity and minimize mean squared error with respect to the observed interval data:

$$\Gamma = \underset{g \in G}{\operatorname{argmin}} \left[ \sum_{k=1}^{K} \frac{i_{k+1} - i_k}{100} \left( \left( \frac{1}{i_{k+1} - i_k} \int_{i_k}^{i_{k+1}} s(i, g) di \right) - \overline{r}_k \right)^2 \right] \tag{IV.3}$$

such that

$$E(y|i) \text{ is weakly increasing in } i. \tag{Monotonicity}$$

The $\frac{i_{k+1}-i_k}{100}$ term weights the mean squared error by the width of each bin.[39]

Any mobility statistic $m$ that is a single-valued function of the CEF, such as interval mobility or the rank-rank gradient, can be defined as $m(\gamma)$.[40] The bounds on $m(\gamma)$ are therefore:

$$\begin{aligned} m^{min} &= \inf\{m(\gamma) \mid \gamma \in \Gamma\} \\ m^{max} &= \sup\{m(\gamma) \mid \gamma \in \Gamma\} \end{aligned} \tag{IV.4}$$

We can further tighten bounds on the CEF by restricting its curvature. We motivate this restriction by the observation that empirical examples of fully supported parent-child rank distributions are relatively smooth (see Figure 5 below), while the analytical bounds on the curvature-unconstrained CEF imply implausible discontinuities, which we highlight in Appendix B. We consider a curvature restriction with the following structure:

$$s(i, \gamma) \text{ is twice-differentiable and } |s''(i, \gamma)| \leq \overline{C}. \tag{Curvature Constraint}$$

This analogous to imposing that the first derivative is Lipshitz.[41] Depending on the value

[39]While we choose to use a weighted mean squared error penalty, in principle $\Gamma$ could use other penalties.

[40]We obtain the rank-rank gradient as the slope parameter which numerically minimizes the following:

$$\underset{\hat{\beta}, \hat{\beta}_0}{\operatorname{argmin}} \Big\{ \int_I E(y|i) - \left( \hat{\beta} i + \hat{\beta}_0 \right) \, di \Big\}.$$

[41]Let $X, Y$ be metric spaces with metrics $d_X, d_Y$ respectively. The function $f : X \to Y$ is *Lipschitz continuous* if there exists $K \geq 0$ such that for all $x_1, x_2 \in X$,

$$d_Y(f(x_1), f(x_2)) \leq K d_X(x_1, x_2).$$

of $\overline{C}$, this constraint may or may not bind.

Note that the most restrictive curvature constraint, $\overline{C} = 0$, is analogous to the assumption of a linear CEF — which will be equivalent to the predicted values from the conventional rank-rank gradient estimation. An intermediate curvature restriction is therefore less restrictive than the default practice in many current studies of mobility.[42]

The results in this section are valid for any conditional expectation function $E(y|x)$, and are not limited to the child rank CEF. One such function that has been of interest to researchers is the conditional expectation of attaining the top quintile (Chetty et al., 2017b). The statistic $\mu_0^{20}$ with this CEF would generate the probability that a child of a bottom quintile parent attains the top quintile. This statistic would otherwise be very difficult to calculate given interval rank data without a boundary at the 20th percentile.

### IV.B Numerical Calculation of Bounds on Mobility

This section describes a method to numerically solve the constrained optimization problem in Equations IV.3 and IV.4. We take a nonparametric approach because observation of CEFs from countries with fully supported parent-child rank distributions do not suggest an obvious parametric form that will fit many distributions (see Figure 5). To make the problem numerically tractable, we solve the discrete problem of identifying the feasible values taken by $E(y|i)$ at each of $N$ discrete values of $i$. In practice, we assume $E(y|i) = s(i, \gamma)$, where $\gamma$ is an $N$-dimensional column vector that defines the value of the CEF between each of the $N$ ranks. We use $N = 100$ in our analysis, corresponding to 100 integer parent ranks, but other values may be useful depending on the application. Given continuity in the latent function, the discretized CEF is a very close approximation of the continuous CEF; increasing the value of $N$ does not change any of our results, but increases computation time significantly.

We solve the problem through a two-step process. Define a $N$-valued vector $\hat{\gamma}$ as a candidate CEF. First, we calculate the minimum MSE from the constrained optimization problem given by Equation IV.3. We then run a second pair of constrained optimization problems

[42] We discuss the selection of $\overline{C}$ in Section IV.D.

that respectively minimize and maximize the value of $m(\hat{\gamma})$, with the additional constraint that the MSE is equal to the value obtained in the first step, denoted $\underline{\text{MSE}}$. Equation IV.5 shows the second stage setup to calculate the lower bound on $m(\hat{\gamma})$:

$$m^{min} = \min_{\hat{\gamma} \in [0,100]^N} m(\hat{\gamma}) \tag{IV.5}$$

such that

$$s(i, \hat{\gamma}) \text{ is weakly increasing in } i \qquad \text{(Monotonicity)}$$

$$|s''(i, \hat{\gamma})| \leq \overline{C}, \qquad \text{(Curvature)}$$

$$\left[ \sum_{k=1}^{K} \frac{\|I_k\|}{100} \left( \left( \frac{1}{\|I_k\|} \sum_{i \in I_k} s(i, \hat{\gamma}) \right) - \overline{r}_k \right)^2 \right] = \underline{\text{MSE}} \qquad \text{(MSE Minimization)}$$

where $I_k$ is the set of discrete values of $i$ between $i_k$ and $i_{k+1}$ and $\|I_k\|$ is the width of bin $k$. The complementary maximization problem obtains the upper bound on $m(\hat{\gamma})$.

Note that setting $m(\gamma) = \gamma_p$ permits us to obtain bounds on the value of the CEF at parent rank $p$. Calculating this for all ranks $p$ from 1 to 100 generates analogous bounds to those derived in proposition 1, but satisfying the additional curvature constraint.

### IV.C Example with Sample Data

In this section, we demonstrate the bounding method using data from India, continuing with the cohort of sons born in the 1960s. Appendix Figure A3 graphs the analytical and numerical upper and lower bounds on $E(y|i)$ at each value of $i$, confirming that the numerical optimization delivers the correct result when $\overline{C} = \infty$. A continuum of functions within these bounds obtain MSE of zero and can be considered equally plausible fits to the data. The bounds on the CEF are considerably tighter at the top of the distribution where the rank bins are relatively narrow. These bounds do not reflect statistical uncertainty; we discuss and present bootstrap confidence sets below.

The four panels of Figure 3 show the effect on the CEF bounds of progressively tightening the curvature constraint, with all bounds calculated numerically. As above, the bounds remain tightest when the parent bin size is small and the function is bounded on both sides. The tighter bounds in Panels C and come at the cost of bias; MSE rises as the curvature constraint makes it impossible to meet all the bin means. Panel D shows the limit case with $\overline{C} = 0$; the CEF in this figure is identical to the predicted values from a regression of son rank on father rank, and is plainly a poor fit to the data.

Since the bin means are measured with sampling error, a confidence set that includes the true bounds with 95% certainty can be calculated via bootstrapping (Imbens and Manski, 2004; Tamer, 2010). We calculate bootstrap confidence sets using 1,000 bootstrap samples from the underlying datasets; these are shown in Figure 4, using the same curvature constraint as Figure 3B.[43] The confidence sets are tight because the census data from which these father and son education values are drawn is very large. In settings with smaller samples or higher variance outcomes, sampling error may be a more important factor. For presentational clarity, we omit bootstrap bounds from further CEF graphs, but they are available from the authors and do not substantively change any of the results described below.

The bounds shown in these graphs represent the envelope of a continuum of CEFs that fit the data equally well. The bounds graphed do not display CEFs themselves. Rather, the bounds on a given scalar mobility statistic reflect the highest and lowest mobility estimate obtained from the set of feasible CEFs.

We present estimates of the rank-rank gradient, $p_{25}$ and $\mu_0^{50}$ in Table 2, for all decadal birth cohorts and under a range of curvature constraints, with 95% bootstrap confidence sets in parentheses. Several patterns emerge. First, in all cases, $\mu_0^{50}$ can be bounded more meaningfully than either $p_{25}$ or the rank-rank gradient. $\mu_0^{50}$ can be calculated with relative precision, as long as there is a parent bin boundary close to the 50$^{\text{th}}$ percentile, a condition that is least well satisfied by the 1950s birth cohort, among whom 60% of fathers report zero

[43]Because the IHDS is weighted, we take weighted random draws that take into account the higher representation of some observations in the population. The bootstrap sample can then be treated as an unweighted sample from the underlying dataset. Because the total value of weights is 10 million, we do not have sufficient computational power to calculate bootstrap samples with the same size as the underlying sample. We chose a bootstrap sample size of 10,000, which allowed us to generate 1,000 bootstrap samples.

years of education. Second, the bounds on all statistics are tighter for more recent birth cohorts. The width of the bounds is largely driven by the size of the bottom parent bin, which gets progressively smaller with rising educational attainment over time.[44]

### IV.D Choosing the Curvature Parameter

The width of the bounds often depends on the curvature threshold $\overline{C}$. Without additional data on the latent rank distribution, it is difficult to estimate an appropriate curvature restriction from the sample being studied. To provide an empirical basis for a reasonable curvature threshold, we therefore examined as many studies as we could find that report uncensored parent-child outcome rank distributions. Where possible, we obtained the underlying data, and where we could not obtain the underlying data we estimated the rank-rank matrix by digitizing the underlying data from the paper's graphics. In each of these studies, we obtained nonparametric fits to the data using 5-piece cubic splines with equally spaced knots.[45] We calculated the maximum absolute value of the 2nd derivative across integer bins in each of these studies.

Figure 5 shows the splines that we fit to each dataset and the corresponding range of the 2nd derivative in each of these functions. Given that many of these studies exhibited significantly more curvature in the tails of the distribution, we also report a second set of maxima from the 10th to the 90th percentile of these distributions. The magnitude of the 2nd derivative is highest in Swedish data, at 0.067 in the full sample and 0.046 in the [10,90] interval. These estimates admittedly may not be representative of the shape of global educational mobility curves, both because they are drawn from income distributions rather than education distributions, and because they are limited to the set of countries which collect and report uncensored matched parent-child income. This said, they are the best available reference for estimation of an appropriate curvature restriction for our context. We propose a conservative

[44]Appendix Table A1 shows the transition matrices that underlie these estimates for each cohort.

[45]Splines with more knots tended to overfit the data (for example, by predicting nonmonotonic segments) and splines with fewer knots often failed to account for curvature in the data, especially in the tails. These estimates are minimally affected by changes in the number of knots.

rule of thumb that limits curvature to approximately twice the upper bound, or $\overline{C} = .1$. This corresponds to the constraint used in Panel B of Figure 3. We use this value going forward, recognizing that it may underestimate the width of the bounds at the tails of the distribution.

### IV.E Simulation: Bounds on Mobility in Denmark

In this section, we simulate our method by taking data from the fully supported parent-child CEF from the Danish income distribution (Boserup et al., 2014), interval-censoring that data, and then recovering bounds on the true CEF from that interval censored data. We begin with data on the average child income rank for every integer parent rank from 1 to 100 (shown in Figure 5B). We interval censor the data according to the same seven bins used in the Indian example above. We then calculate bounds on the CEF using only the binned data.

Panels A–D of Figure 6 show the result of this exercise, with curvature limits corresponding to Panels A–D of Figure 3. The dashed lines show the underlying uncensored spline approximation to the parent-child distribution. The solid circles show the constructed bin means of the censored data; these are the only data that we use for the optimization. The solid lines show the upper and lower envelopes that we calculate for the nonparametric CEF. As above, the bounds are tighter at the top of the distribution where we have artificially imposed smaller bins, and wider at the bottom. Compared with the Indian exercise, the bounds are tighter across the entire distribution because of the higher level of mobility in Denmark: since the expected ranks of children are more similar across parents' ranks, the monotonicity assumption significantly constrains the shape of the CEF. In Panels A through C, the true data are almost entirely contained by the bounds. It is important to note that the true CEF is not always centered by the bounds; from $p_{30}$ to $p_{40}$ the true CEF is near the upper bound, and from $p_1$ to $p_{10}$ it is nearer the lower bound. The linear Panel D significantly overestimates mobility at the bottom of the distribution.

Almost any parametric best fit to the censored data would miss the concavity at the lower tail of the distribution. The bias of parametric methods is increasing in the non-linearity of the underlying CEF and the bin sizes. The strength of our method is that in the domains where interval censoring is most severe, it makes clear that we can at best obtain bounds on the partially identified CEF.

### IV.F Interval Censoring in the Child Distribution

Thus far, we have focused on interval censoring in the parent rank distribution and defined $\overline{r}_k$ by assuming that each child has the midpoint rank in his or her education bin. This is equivalent to assuming, for example, that the expected latent rank of a primary-educated child born to a parent with no education is equal to the expected latent rank of a primary-educated child born to a parent with a middle school education. But if the latent rank of children *within* the outcome bins where they are observed is positively correlated with the ranks of their parents, then estimates of intergenerational mobility will be biased downward. This concern is not unique to our study; rather it is the complementary problem to the interval censoring of parent rank that we study above.

In principle, we could use the same constrained optimization approach to bound the entire distribution of child ranks conditional on parent ranks, rather than just the conditional expectation of child rank. A numerical approach to this problem might bound each cell of the 100x100 integer rank transition matrix, taking as sample analogs the 49 means observed in the 7x7 transition matrix. Parameterized as above, this would be a problem with 49 moment conditions, 10,000 parameters, and at least 10,000 constraints. While there may be simplifications that can make this problem computationally tractable, it is beyond the scope of this study, though we explore the problem in more detail in Appendix C.

In addition to being more difficult computationally, this problem is also a less important source of bias for three reasons. First, coarse rank data poses the greatest problem for mobility estimation when the top or bottom bin is large. Since almost all countries are becoming better off over time, the lowest rank bin in the child distribution is nearly always smaller than the lowest rank bin for parents.[46] This argument may hold less relevance in richer

[46] In India, for sons born in the 1980s, 11% report zero education, compared with 37% of their fathers. For the 1950s birth cohort, 31% of sons report zero education, compared to 60% of fathers. The worst case in the son distribution is thus comparable to the best case in the father distribution.

countries where education may be increasingly topcoded over time, but researchers in these contexts are less often required by data availability to focus on education.

Second, data on children are more widely available than data on parents. Matched parent-child data often contain many characteristics of children, but only a few coarsely measured characteristics of parents. Rank-rank distributions can be estimated and compared with different underlying outcomes on the $X$ and $Y$ axes, so there are many contexts where interval censoring will be a concern for the parent generation but not for the child generation. For example, to measure children's educational outcomes when a large number of children have completed college, Chetty et al. (2014a) resolve the coarse data problem by ranking children according to their college quality rank. They show that this variable is highly monotonic in parent income rank, which in passing further supports our assumption of monotonic latent education ranks.

Third, the availability of additional child data also makes it possible to use other variables to estimate the relationship between parent rank and latent child rank *within* each parent-child education rank bin. For example, we can use child wages within education bins to measure the extent to which child wages are sorted on parent ranks *within* interval censored child education bins. In contrast, this approach does not help us describe the latent parent rank distribution without additional data on parents.

In Appendix C, we explore several of these arguments in more detail. We derive a numerical solution to the intermediate problem of bounding the child CEF when child data are interval censored but parent data are observed precisely. We show that in India, the mobility bounds when son data are censored are considerably tighter than when parent data are censored, confirming that censoring of parent rank data is the more substantive concern. Second, we predict latent parent and child ranks *within* parent bins of the education transition matrix by using sons' wages. The result suggests that our assumption that son rank is equal to the rank bin midpoint is a close approximation to the empirical distribution. These results strongly suggest that taking censoring of son rank data into account will only minimally affect our estimates below. The more computationally difficult problem of simultaneously addressing coarseness in both parent and child rank distributions is left for future work.

## V Application: Changes in Intergenerational Mobility in India

In this section, we calculate bounds on intergenerational educational mobility in India. There are three main findings. First, we show that conventional estimates of rank-based mobility measures provide limited information about changes over time and across groups once the coarseness of rank groups is taken into account. However, our new measure, upward interval mobility can be estimated with much tighter bounds. Second, using this measure of upward mobility, we show that population intergenerational mobility in India is static over time and is considerably lower than in the United States. Third, India's historically marginalized Scheduled Castes and Schedules Tribes have made substantial mobility gains since the 1960s, closing the upward mobility gap by about a third. These gains are almost exactly reversed for the Muslim population, who have experienced a significant loss in mobility since the 1960s.

All mobility estimates and curves are calculated using the curvature restriction proposed in Section IV.D. Alternate curvature parameters lead to largely similar conclusions.

### V.A Indian Intergenerational Mobility Over Time

To examine intergenerational mobility over time in India, we compare the joint distribution of father and son ranks for an older cohort, born in the 1950s and a younger cohort born in the 1980s. Panel A of Figure 7 displays the raw data that are available to the researcher: for each father education rank bin, we observe the mean son rank in the outcome distribution. Since we do not observe father education ranks within education bins, we plot father ranks at the median rank in each interval where education is reported. The solid vertical line shows the bin boundary for the least educated fathers in the older cohort; 60% of fathers are in this bin. The dashed line shows the equivalent bin boundary for fathers in the younger cohort; 38% report fewer than two years of education.[47]

[47] Appendix Table A1 presents the full transition matrices for all four decadal cohorts in our sample, from which these nonparametric plots are derived.

Panel B of Figure 7 shows bounds on the latent father-son rank distribution.[48] The large share of fathers with zero education highlights the challenge of describing changes in intergenerational mobility at the bottom of the distribution. The CEF for the younger generation has tighter bounds because the lowest parent bin is smaller. These bounds are contained almost entirely within the bounds for the older generation. The bounds on the CEF at any point in the bottom half of the distribution are consistent with both zero change over time and with substantial mobility gains. However, we can identify a small decrease in the intergenerational persistence of the highest positions in the education distribution.[49]

Figure 8 shows bounds on the three mobility statistics discussed above for each decadal cohort: the rank-rank gradient, absolute upward mobility ($p_{25}$), and upward interval mobility ($\mu_0^{50}$). For reference, we plot recent estimates of similar mobility measures from USA and Denmark.[50] The graph highlights the advantage of upward interval mobility in a context with significant interval censoring. The rank-rank gradient and absolute upward mobility provide little information on how Indian mobility has changed over time; the estimates are consistent with both large mobility gains and with no change in mobility over time. In contrast, upward interval mobility is estimated with tight bounds in all periods. According to this measure, upward mobility has changed very little over the four decades studied; there is a small gain from the 1950s to the 1960s, followed by a small decline from the 1960s to the 1980s. On average, Indian mobility is the same distance below the United States as mobility in the United States is below Denmark. Table 2 reports the bounds for each measure and cohort with bootstrap confidence sets under a range of curvature restrictions. The choice of curvature does not substantively change these conclusions, though estimates with less

[48] Given multiple bounds on a single graph, we have omitted bootstrap standard errors for legibility. As demonstrated in Figure 4, these imply a slight widening of the bounds across the distribution and do not change our conclusions.

[49] Appendix Figures A4 and A5 show similar estimates with the old cohort defined respectively as born in the 1960s and in the 1970s. Our results do not change substantially when considering the intermediate cohorts; we present only the 1950s and 1980s birth cohorts here for graph legibility.

[50] Rank-rank correlations of education are from Hertz (2008). For absolute mobility, we calculate $p_{25}$ for the U.S. and Denmark from the distributions shown in Figure 5, with data from Chetty et al. (2014a).

restrictive curvature are necessarily less precise.[51]

### V.B Intergenerational Mobility of Marginalized Groups

In this section, we decompose the secular stagnation in intergenerational mobility into subgroup estimates for Scheduled Castes and Scheduled Tribes (SC/STs), Muslims, and the remainder of the population. Because the SECC does not report religion, all these estimates are derived from the IHDS alone (with both coresident and resident fathers and sons), and thus have slightly wider confidence sets than the estimates above.[52]

Panels A and B of Figure 9 show bounds on the child rank CEF given parent rank for SC/STs, Muslims and all others, respectively for sons born in the 1950s and in the 1980s respectively.[53] We rank fathers and sons according to their position in the national education distribution, because we are interested in the ability of marginalized groups to move up in the national distribution, rather than relative to other members of their group. In the 1950s birth cohorts, SC/STs and Muslims are virtually identical in status, but by the 1980s, SC/STs have moved ahead. This change would be somewhat obscured by looking only at mean education, as expected child ranks for SC/STs and Muslims in the top half of the parent rank distribution have remained similar, with the possible exception of greater SC/ST persistence of the very highest ranks.

In Panel C of Figure 9 we present statistics for absolute upward mobility ($p_{25}$) and for upward interval mobility ($\mu_0^{50}$), by cohort and subgroups.[54] As above, the bounds on $p_{25}$ are

[51] Appendix Figure A6 shows comparable estimates using only the IHDS 2011, which are substantively unchanged. Appendix A7 shows estimates for the same cohorts using the IHDS 2005 to gauge whether survivorship bias influences results. The results suggest moderate survivorship bias in mobility estimates for the 1950-59 cohort, but no bias for the more recent cohorts. The patterns of interest described in this paper largely begin with the 1960s birth cohort, and thus the conclusions are unchanged by this bias in the 1950-59 cohort. If anything, the survivor bias adjusted results suggest the trend from the 1960s to the 1980s in fact began in the 1950s.

[52] Conducting this analysis requires the assumption that the latent rank distributions of each subgroup are comparable within each education bin, an assumption implicitly made in other studies comparing educational outcomes across groups.

[53] Appendix Figure A8 shows just the rank bin means for each group.

[54] We do not examine the rank-rank gradient separately for population subgroups, because this measure describes how privileged members of a subgroup do relative to other members of the same subgroup. Our interest in this paper is in documenting how members of these historically marginalized groups perform as compared with the entire population.

too wide to be informative, while estimates of $\mu_0^{50}$ are much tighter.

We note two features in this figure. First, once SC/STs and Muslims are treated separately, upward mobility for the remainder of the population is now comparable to that of the United States, and has been since the 1950s. Low intergenerational mobility in India can thus be significantly explained by low mobility in these two minority groups. Second, the intergenerational mobility of SC/STs and Muslims has diverged dramatically since the 1960s. In the 1960s, the expected outcomes of SC/ST and Muslim children in the bottom half of the distribution were comparable to each other: SC/STs could expect to attain the 33rd percentile, while Muslims could expect to attain the 31st. By the 1980s, SC/STs in the bottom half of the distribution were attaining the 36th percentile on average, closing more than half the gap between themselves and the non-Muslim, non-SC/ST population over this period. In contrast, a child born in the 1980s to a Muslim family attain on average only the 29th percentile in the outcome distribution. We have focused on changes from the 1960s to the present as the numbers from the 1950s are subject to survivorship bias. However, beginning the analysis in the 1950s does not change the conclusion of significant divergence between SC/STs and Muslims.

We are aware of no previous study on India that has examined the ability of subgroups to improve their relative ranks in the entire population, exactly because it is difficult to generate comparable rank estimates across cohorts.

## VI Conclusion

Our work highlights a key challenge in measuring educational and occupational mobility: when rank data are reported in coarse bins, the parent-child relationship within bins is important for mobility but is unobservable to the researcher. Using a small set of structural assumptions, we show how estimates of intergenerational mobility can be bounded to take into account the parts of the distribution that the researcher does not observe.

The key insight in our analysis is that we can meaningfully bound the CEF with just the monotonicity and constrained curvature assumptions. Once we obtain bounds on the values taken by the CEF, we can bound any mobility statistic that is a function of the CEF. The most limiting curvature restraint generates the rank-rank gradient, making our method a generalization of one of the most widely used mobility statistics.

We also show that a new group of interval mobility measures can be bounded tightly even under conditions of severe interval censoring. Bounds on interval mobility can be generated analytically from interval rank data, providing a simple tool for mobility researchers to compare expected outcomes in the bottom half of the distribution even when no rank boundary explicitly defines the bottom half. This measure is most useful in contexts where the bottom rank bins are very large, as is often the case in developing countries and in older generations.

Using data from India, we show that education-based estimates of the father-son rank-rank gradient have considerably less precision than previously thought, once interval censoring of rank data is taken into account. Using our new measure, we show that upward mobility in India has barely changed from the 1950s to the 1980s. This lack of change can be decomposed into substantial gains for SC/STs and substantial losses for Muslims. The latter result has not previously been noted in part because there has previously been no methodology for creating comparable rank bins across cohorts.

Although we have focused on an application to mobility, we have also proved new nonparametric bounds on conditional expectation functions when the distribution of the variables are known. This result has immediate applications in the health literature, where consistent estimation across changing rank bins is frequently desired, and may be useful in any context with interval data that follows a known distribution.

## References


**Asher, Sam and Paul Novosad**, "Rural Roads and Structural Transformation," 2017. Working paper.

**Aydemir, Abdurrahman and Hakki Yazici**, "Intergenerational Education Mobility and the Level of Development: Evidence from Turkey," 2016.

**Azam, Mehtabul and Vipul Bhatt**, "Like Father, Like Son? Intergenerational Educational Mobility in India," *Demography*, 2015, *52* (6).

**Bagde, Surendrakumar, Dennis Epple, and Lowell Taylor**, "Does affirmative action work? Caste, gender, college quality, and academic success in India," *American Economic Review*, 2016, *106* (6).

**Bertrand, Marianne, Rema Hanna, and Sendhil Mullainathan**, "Affirmative action in education: Evidence from engineering college admissions in India," *Journal of Public Economics*, 2010, *94* (1-2), 16–29.

**Björklund, Anders, Mikael Lindahl, and Erik Plug**, "The Origins of Intergenerational Associations: Lessons from Swedish Adoption Data," *Quarterly Journal of Economics*, 2006, *121* (3), 999–1028.

**Black, S, P Devereux, and Kjell G Salvanes**, "Why the Apple Doesn't Fall: Understanding Intergenerational Transmission of Human Capital," *American Economic Review*, 2005, *95* (1).

**Black, Sandra E. and Paul J. Devereux**, "Recent Developments in Intergenerational Mobility," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Amsterdam: North Holland Press, 2011.

**Boserup, Simon Halphen, Wojciech Kopczuk, and Claus Thustrup Kreiner**, "Stability and persistence of intergenerational wealth formation: Evidence from Danish wealth records of three generations," 2014.

**Bound, John, Arline T Geronimus, Javier M Rodriguez, and Timothy A Waidmann**, "Measuring Recent Apparent Declines In Longevity: The Role of Increasing Educational Attainment," *Health Affairs*, 2015, *34* (12).

**Bratberg, Espen, Jonathan Davis, Bhashkar Mazumder, Martin Nybom, Daniel Schnitzlein, and Kjell Vaage**, "A Comparison of Intergenerational Mobility Curves in Germany, Norway, Sweden and the U.S.," *The Scandinavian Journal of Economics*, 2015, *119* (1).

**Bratsberg, Bernt, Knut Røed, Oddbjørn Raaum, Robin Naylor, Markus Jäntti, Tor Eriksson, and Eva Österbacka**, "Nonlinearities in intergenerational earnings mobility: Consequences for cross-country comparisons," *Economic Journal*, 2007, *117* (519), C72–C92.

**Case, Anne and Angus Deaton**, "Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century," *Proceedings of the National Academy of Sciences*, 2015, *112* (49).

**___ and ___** , "Mortality and morbidity in the 21st century," 2017. Working Paper.

**Chetty, Raj, David Grusky, Maximilian Hell, Nathaniel Hendren, Robert Manduca, and Jimmy Narang**, "The fading American dream: trends in absolute income mobility since 1940," *Science*, 2017, *356* (6336).

**___ , John N Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan**, "Mobility Report Cards: The Role of Colleges in Intergenerational Mobility," 2017. NBER Working Paper No. 23618.

**___ , Nathaniel Hendren, Patrick Kline, and Emmanuel Saez**, "Where is the land of opportunity? The geography of intergenerational mobility in the United States," *Quarterly Journal of Economics*, 2014, *129* (4), 1553–1623.

———, ———, ———, ———, **and Nicholas Turner**, "Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility," *American Economic Review: Papers & Proceedings*, 2014, *104* (5), 141–147.

**Clark, Gregory**, "What is the True Rate of Social Mobility in Sweden? A Surname Analysis, 1700-2012," 2012. Working paper.

———, *The Son Also Rises: Surnames and the History of Social Mobility*, Princeton University Press, 2014.

**Corak, Miles**, "Income Inequality, Equality of Opportunity, and Intergenerational Mobility," *Journal of Economic Perspectives*, 2013, *27* (3).

**Dardanoni, Valentino, Mario Fiorini, and Antonio Forcina**, "Stochastic Monotonicity in Intergenerational Mobility Tables," *Journal of Applied Econometrics*, 2012, *27.*

**Dreze, Jean and Amartya Sen**, *An Uncertain Glory: India and its Contradictions*, Oxford University Press, 2013.

**Dunn, Christopher E.**, "Intergenerational Economic Mobility around the World," *B.E. Journal of Economic Analysis and Policy*, 2007, *7* (2), 1–40.

**Emran, M.S. and F. Shilpi**, "Intergenerational Occupational Mobility in Rural Economy: Evidence from Nepal and Vietnam," *Journal of Human Resources*, 2011, *46* (2).

——— **and Forhad Shilpi**, "Gender, Geography, and Generations: Intergenerational Educational Mobility in Post-Reform India," *World Development*, 2015, *72.*

———, **W. Greene, and F. Shilpi**, "When Measure Matters: Coresidency, Truncation Bias, and Intergenerational Mobility in Developing Countries," *Journal of Human Resources (forthcoming)*, 2017.

**Field, Erica, Seema Jayachandran, and Rohini Pande**, "Do Traditional Institutions Constrain Female Entrepreneurship? A Field Experiment on Business Training in India," *American Economic Review: Papers & Proceedings*, 2016, *100* (2).

**Frisancho Robles, Veronica C. and Kala Krishna**, "Affirmative Action in Higher Education in India: Targeting, Catch Up and Mismatch," 2012. NBER Working Paper No. 17727.

**Güell, Maia, José V Rodríguez Mora, and Christopher I. Telmer**, "The informational content of surnames, the evolution of intergenerational mobility, and assortative mating," *Review of Economic Studies*, 2013, *82* (2).

**Guest, Avery M, Nancy S Landale, and James C Mccann**, "Intergenerational Occupational Mobility in the Late 19th Century United States," *Social Forces*, 1989, *68* (2).

**Hertz, Tom**, "A group-specific measure of intergenerational persistence," *Economics Letters*, 2008, *100* (3), 415–417.

**Hilger, Nathaniel G**, "The Great Escape: Intergenerational Mobility Since 1940," 2016. NBER Working Paper No. 21217.

**Hnatkovska, Viktoria, Amartya Lahiri, and Sourabh B. Paul**, "Castes and Labor Mobility," *American Economic Journal: Applied Economics*, 2012, *4* (2).

———, ———, **and** ———, "Breaking the caste barrier: intergenerational mobility in India," *The Journal of Human Resources*, 2013, *48* (2).

**Ho, Kate and Adam M. Rosen**, "Partial Identification in Applied Research: Benefits and Challenges," 2015. Working paper.

**Imbens, Guido W. and Charles F. Manski**, "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 2004, *72* (6).

**Ito, Takahiro**, "Caste discrimination and transaction costs in the labor market: Evidence from rural North India," *Journal of Development Economics*, 2009, *88* (2), 292–300.

**Knight, John, Terry Sicular, and Ximing Yue**, "Educational Inequality in China: The Intergenerational Dimension," 2011.

**Lindahl, Mikael, Mårten Palme, Sofia Sandgren Massih, and Anna Sjögren**, "Long-Term Intergenerational Persistence of Human Capital: An Empirical Analysis of Four Generations," *Journal of Human Resources*, 2012, *50* (1).

**Long, Jason and Joseph Ferrie**, "Intergenerational Occupational Mobility in Great Britain and the United States Since 1850," *American Economic Review*, 2013, *103* (4).

**Maitra, Pushkar and Anurag Sharma**, "Parents and Children: Education Across Generations in India," 2009. Working Paper.

**Manski, Charles**, *Partial Identification of Probability Distributions*, Springer-Verlag, 2003.

**Manski, Charles F. and Elie Tamer**, "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 2002, *70* (2).

**Mohammed, A R Shariq**, "Does A Good Father Now Have To Be Rich? Intergenerational Income Mobility in Rural India," 2016. Working paper.

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

**Piraino, Patrizio**, "Intergenerational Earnings Mobility and Equality of Opportunity in South Africa," *World Development*, 2015, *67* (March), 396–405.

**Restuccia, Diego and Carlos Urrutia**, "Intergenerational Persistence of Earnings: The Role of Early and College Education," *American Economic Review*, 2004, *94* (5).

**Roemer, John.**, "Equality of Opportunity: Theory and Measurement," *Journal of Economic Literature*, 2016, *54* (4).

**Sachar Committee Report**, "Social, Economic and Educational Status of the Muslim Community of India," Technical Report, Government of India 2006.

**Solon, Gary**, "Intergenerational Mobility in the Labor Market," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Amsterdam: North Holland Press, 1999, pp. 1761–1800.

**Tamer, Elie**, "Partial Identification in Econometrics," *Annual Review of Economics*, 2010, *2* (1), 167–195.

**The Economist**, "An Uncertain Community," Oct 29, 2016.

**Wantchekon, Leonard, Marko Klašnja, and Natalija Novta**, "Education and Human Capital Externalities: Evidence from Benin," *The Quarterly Journal of Economics*, 2015, *130* (2), 703–757.

**Xie, Yu and Alexandra Killewald**, "Intergenerational occupational mobility in Great Britain and the United States since 1850: Comment," *The American Economic Review*, 2013, *103* (5).

**Table 1**
Bin Sizes in Studies of Intergenerational Mobility

| Study | Country | Birth Cohort of Son | Number of Parent Outcome Bins | Population Share in Largest Bin |
|---|---|---|---|---|
| Aydemir and Yazici (2016) | Turkey | 1990[55] | 15 | 39% |
| | Turkey | 1960[56] | 15 | 78% |
| Dunn (2007) | Brazil | 1972–1981 | $> 18$ | 20%[57] |
| Emran and Shilpi (2011) | Nepal, Vietnam | 1992-1995 | 2 | 83% |
| Güell et al. (2013) | Spain | $\sim$ 2001 | 9 | 27%[58] |
| Guest et al. (1989) | USA | $\sim$ 1880 | 7 | 53.2% |
| Hnatkovska et al. (2013) | India | 1918-1988 | 5 | Not reported |
| Knight et al. (2011) | China | 1930–1984 | 5 | 29%[59] |
| Lindahl et al. (2012) | Sweden | 1865-2005 | 8 | 34.5% |
| Long and Ferrie (2013) | Britain | $\sim$ 1850 | 4 | 57.6% |
| | Britain | $\sim$ 1949-55 | 4 | 54.2% |
| | USA | $\sim$ 1850-51 | 4 | 50.9% |
| | USA | $\sim$ 1949-55 | 4 | 48.3% |
| Piraino (2015) | South Africa | 1964–1994 | 6 | 36% |

Table 1 presents a review of other papers analyzing educational and occupational mobility. The sample is not representative: we focus on other papers where interval censoring may be a concern. The column indicating number of parent outcome bins refers to the number of categories for the parent outcome used in the main specification. The outcome is education in all studies with the exception of Long and Ferrie (2013) and Guest et al. (1989), where the outcome is occupation.

[56] Includes all people born after about 1990.

[57] Includes all people born after about 1960.

[58] This is the proportion of sons in 1976 who had not completed one year of education — an estimate of the proportion of fathers in 2002 with no education, which is not reported.

[59] Estimate is from the full population rather than just fathers.

[60] This reported estimate does not incorporate sampling weights; estimates with weights are not reported.

**Table 2**

Bounds on Intergenerational Educational Mobility in India

| Cohort | Gradient ($\overline{C} = \infty$) | $p_{25}$ ($\overline{C} = \infty$) | $\mu_0^{50}$ ($\overline{C} = \infty$) | Gradient ($\overline{C} = 0.20$) | $p_{25}$ ($\overline{C} = 0.20$) | $\mu_0^{50}$ ($\overline{C} = 0.20$) |
|---|---|---|---|---|---|---|
| 1950-59 | [0.457, 0.742] | [13.0, 58.3] | [34.8, 38.7] | [0.474, 0.722] | [25.5, 48.1] | [34.8, 37.9] |
| | (0.447, 0.763) | (10.2, 59.8) | (34.0, 39.0) | (0.464, 0.745) | (24.2, 48.3) | (34.0, 38.4) |
| 1960-69 | [0.436, 0.655] | [22.2, 54.5] | [37.0, 39.1] | [0.444, 0.639] | [29.3, 49.3] | [37.0, 38.8] |
| | (0.421, 0.677) | (19.7, 54.8) | (36.3, 39.5) | (0.429, 0.661) | (28.0, 49.5) | (36.3, 39.3) |
| 1970-79 | [0.463, 0.595] | [29.0, 48.6] | [37.8, 37.8] | [0.468, 0.584] | [32.2, 48.3] | [37.8, 37.8] |
| | (0.455, 0.616) | (26.8, 49.7) | (37.3, 38.0) | (0.461, 0.603) | (31.9, 49.1) | (37.3, 38.0) |
| 1980-89 | [0.500, 0.565] | [32.3, 42.3] | [36.8, 36.8] | [0.505, 0.556] | [33.3, 42.8] | [36.8, 36.8] |
| | (0.488, 0.591) | (30.2, 43.6) | (36.4, 37.3) | (0.492, 0.582) | (32.8, 43.6) | (36.4, 37.3) |

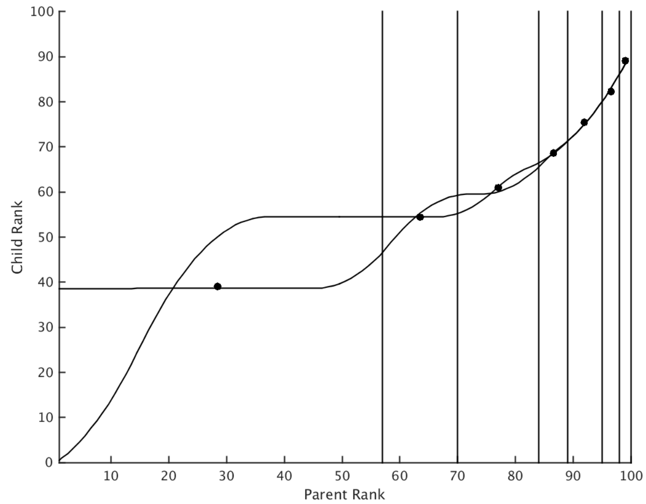
| Cohort | Gradient ($\overline{C} = 0.10$) | $p_{25}$ ($\overline{C} = 0.10$) | $\mu_0^{50}$ ($\overline{C} = 0.10$) | Gradient ($\overline{C} = 0$) | $p_{25}$ ($\overline{C} = 0$) | $\mu_0^{50}$ ($\overline{C} = 0$) |
|---|---|---|---|---|---|---|
| 1950-59 | [0.492, 0.702] | [28.4, 44.9] | [34.9, 37.1] | 0.587 | 35.6 | 35.6 |
| | (0.480, 0.727) | (27.2, 45.3) | (34.0, 37.8) | (0.577, 0.595) | (35.4, 35.9) | (35.4, 35.9) |
| 1960-69 | [0.452, 0.629] | [31.2, 46.3] | [37.0, 38.5] | 0.538 | 36.8 | 36.8 |
| | (0.436, 0.629) | (31.1, 46.5) | (36.9, 38.9) | (0.530, 0.553) | (36.5, 37.0) | (36.5, 37.0) |
| 1970-79 | [0.472, 0.577] | [33.4, 46.2] | [37.8, 37.8] | 0.534 | 36.9 | 36.9 |
| | (0.465, 0.597) | (32.7, 46.5) | (37.3, 38.0) | (0.524, 0.549) | (36.6, 37.2) | (36.6, 37.2) |
| 1980-89 | [0.506, 0.548] | [33.8, 42.4] | [36.8, 36.8] | 0.537 | 36.8 | 36.8 |
| | (0.494, 0.575) | (33.3, 43.1) | (36.3, 37.3) | (0.523, 0.551) | (36.5, 37.2) | (36.5, 37.2) |

The table shows estimates of bounds on three scalar mobility statistics, for different decadal cohorts and under different restrictions $\overline{C}$ on the curvature of the child rank conditional expectation function given parent rank. The rank-rank gradient is the slope coefficient from a regression of son education rank on father education rank. $p_{25}$ is absolute upward mobility, which is the expected rank of a son born to a family at the 25th percentile. $\mu_0^{50}$ is upward interval mobility, which is the expected rank of a son born below to a family below the 50th percentile. When $\overline{C} = 0$, the bounds shrink to point estimates. Bootstrap 95% confidence sets are displayed in parentheses below each estimate based on 1000 bootstrap samples.

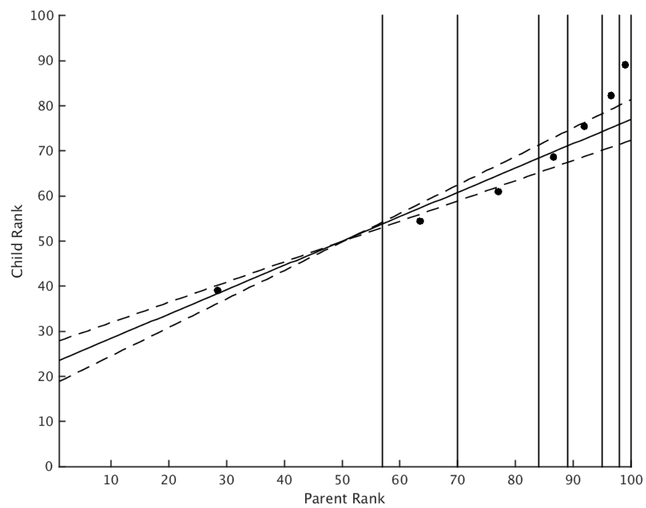
# **Figure 1**

Joint Distribution of Father-Son Education Ranks
and Sample Conditional Expectation Functions

Panel A




Panel B




---

Figure 1 presents example conditional expectation functions of child education ranks given parent education ranks. The sample is Indian sons born between 1960 and 1969 and their fathers. The circles represent the mean ranks of fathers and sons in each father bin in the education distribution. The vertical lines show the boundaries between parent education rank bins. In Panel A, the solid lines show two different conditional expectation functions that minimize the MSE with respect to the bin means. Panel B shows linear approximations to the two functions in Panel A (dashed lines) as well the predicted values from a regression of son bin means on father bin means, weighted by father bin size (solid line).

**Figure 2**

Analytical Bounds on the Child Rank CEF

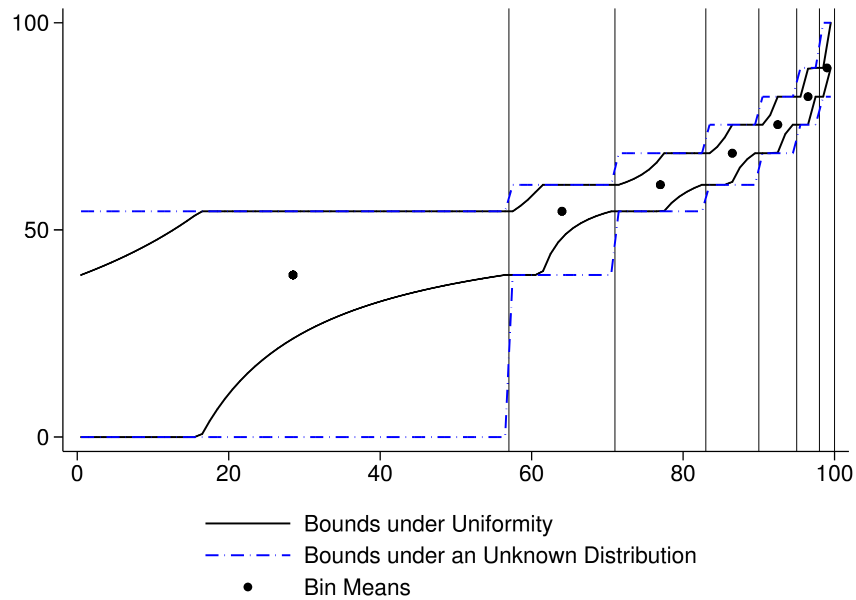



Figure 2 shows bounds on the conditional expectation of child rank given parent rank. The vertical lines show the bin boundaries and the points show the mean value of child rank in each bin. The points are centered at the rank bin medians on the X axis. The dashed lines show analytical bounds when the distribution of the $X$ variable is unknown (Manski and Tamer, 2002). The solid line shows analytical bounds when the distribution of the $X$ variable is uniform.

**Figure 3**

Father-Son Rank Conditional Expectation Function Bounds
Under Curvature Constraints

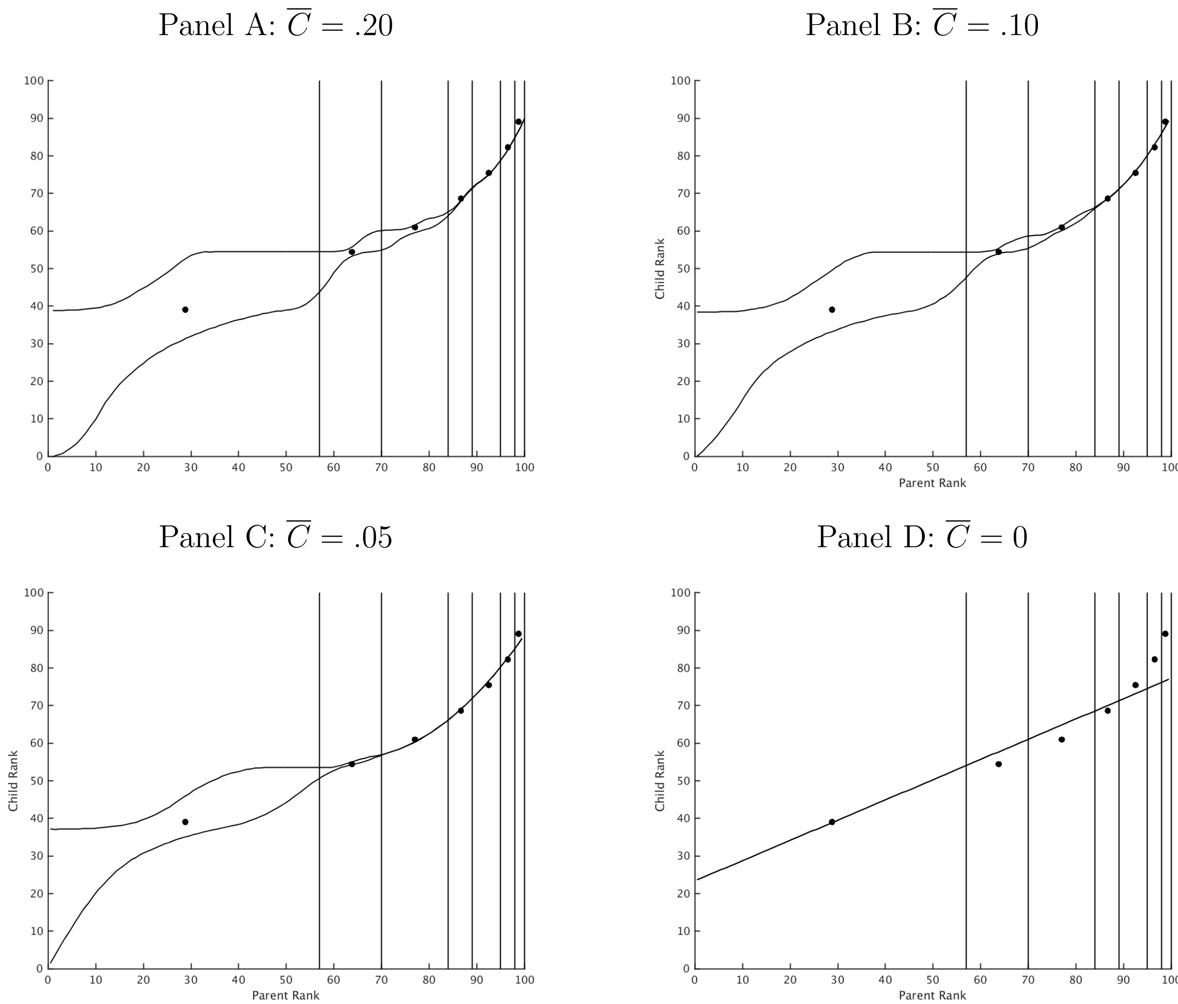



Figure 3 presents bounds on conditional expectation functions of child education ranks given parent education ranks for Indian men born in the 1960s and their fathers. The functions in each panel represent the upper and lower bounds on the CEF at each parent rank, obtained under different curvature restrictions. All four panels impose monotonicity, as well as the constraint on the second derivative given in the panel title. The circles represent the mean ranks of fathers and sons in each father bin in the education distribution. The vertical lines show the boundaries between parent education rank bins.

**Figure 4**

Bounds on Conditional Expectation of Child Rank:
Bootstrap Standard Errors

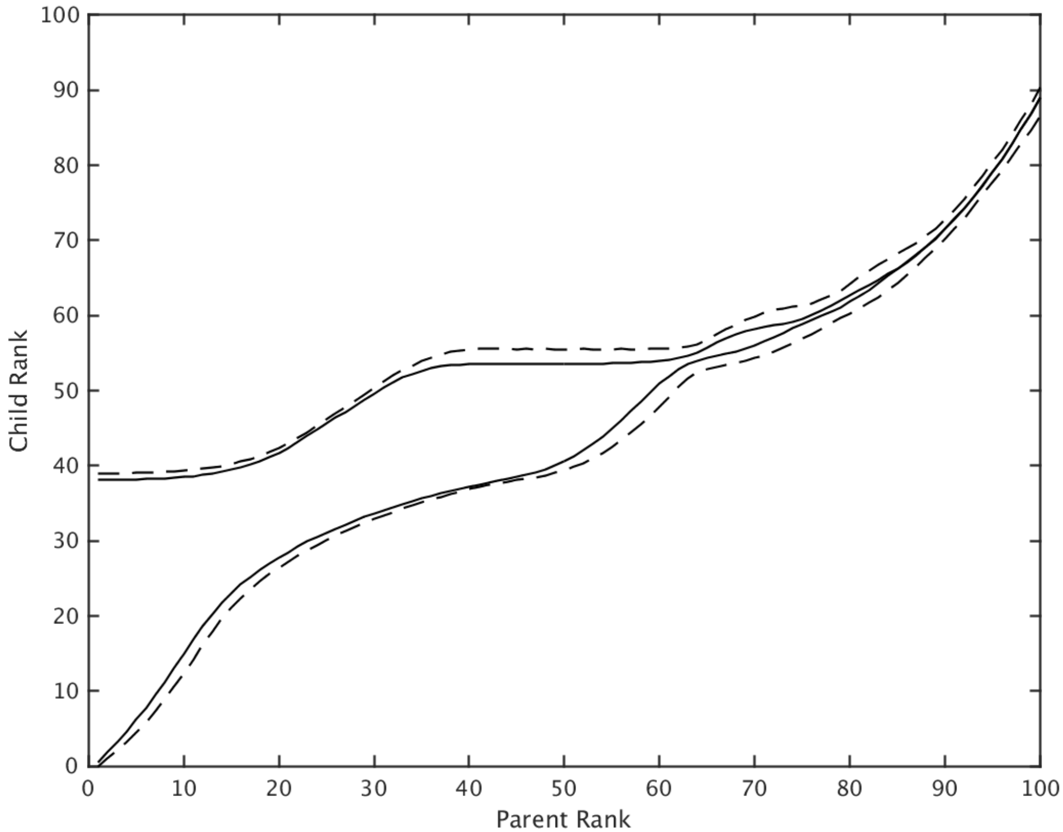



Figure 4 presents upper and lower bounds on the conditional expectation function of child education rank given parent education rank, as well as the bootstrapped 95% confidence set of these bounds, for Indian men born in the 1960s and their fathers. The bounds obtained require monotonicity and that the numerical second derivative of the CEF be less than or equal to .10. We take 1000 bootstraps of the moments and fit bounds to each of these moments. The solid lines present the mean bound obtained across bootstraps. The dashed lines represent the 95% confidence set. We do not display bin boundaries and points because they vary across each bootstrap draw.

## Figure 5

Spline Approximations to Empirical Rank Distributions in the Literature

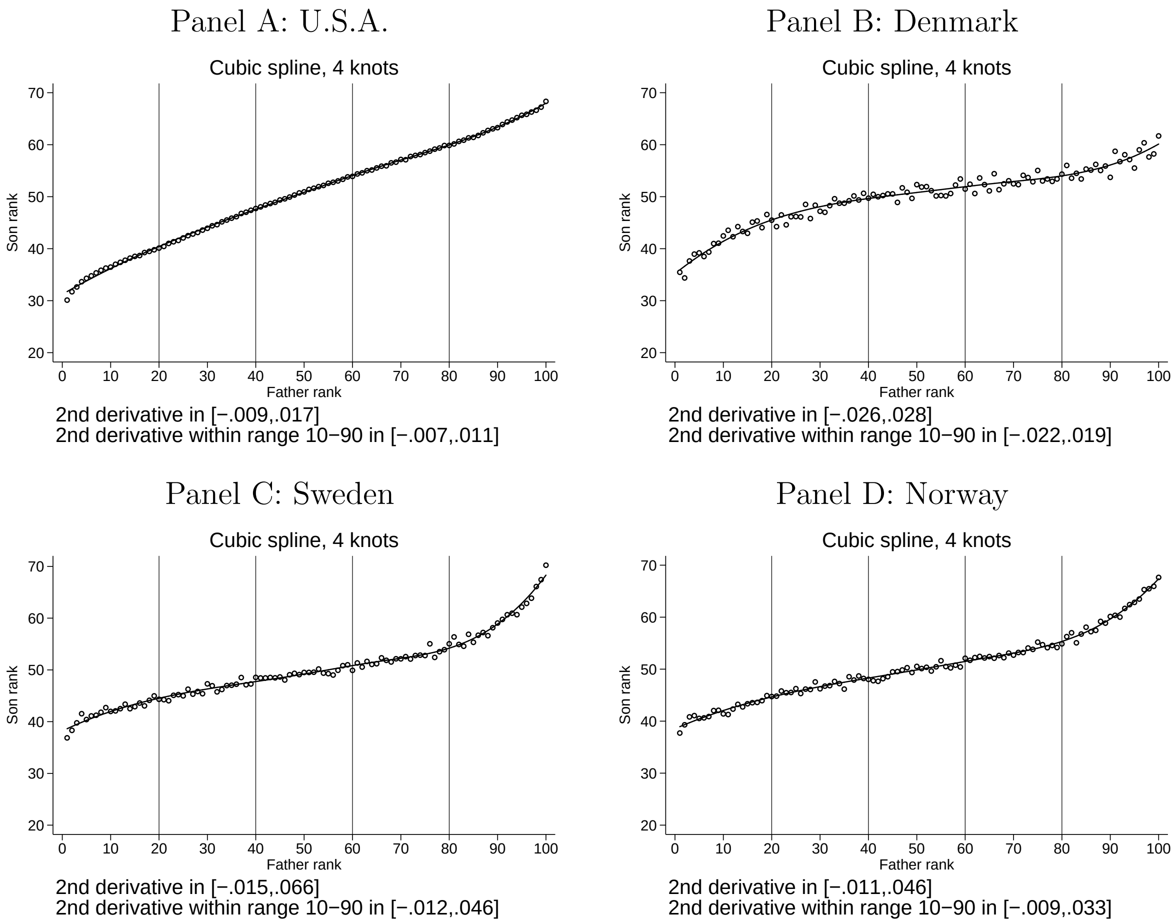



Figure 5 presents estimates of the conditional expectation functions obtained from fully supported parent-child rank-rank income distribution in several developed countries. The data for U.S.A. and Denmark come from Chetty et al. (2014a), who obtained the Denmark data from Boserup et al. (2014). The data for Sweden and Norway come from Bratberg et al. (2015). The CEFs were fitted using cubic splines, with knots at 20, 40, 60, and 80 (as indicated by the vertical lines). The functions plot the best cubic spline fit to each series, and the circles plot the underlying data. Underneath each graph, we present the range of the second derivative across the support and in the domain 10–90.

# Figure 6

Nonparametric Bounds using Danish Mobility Data

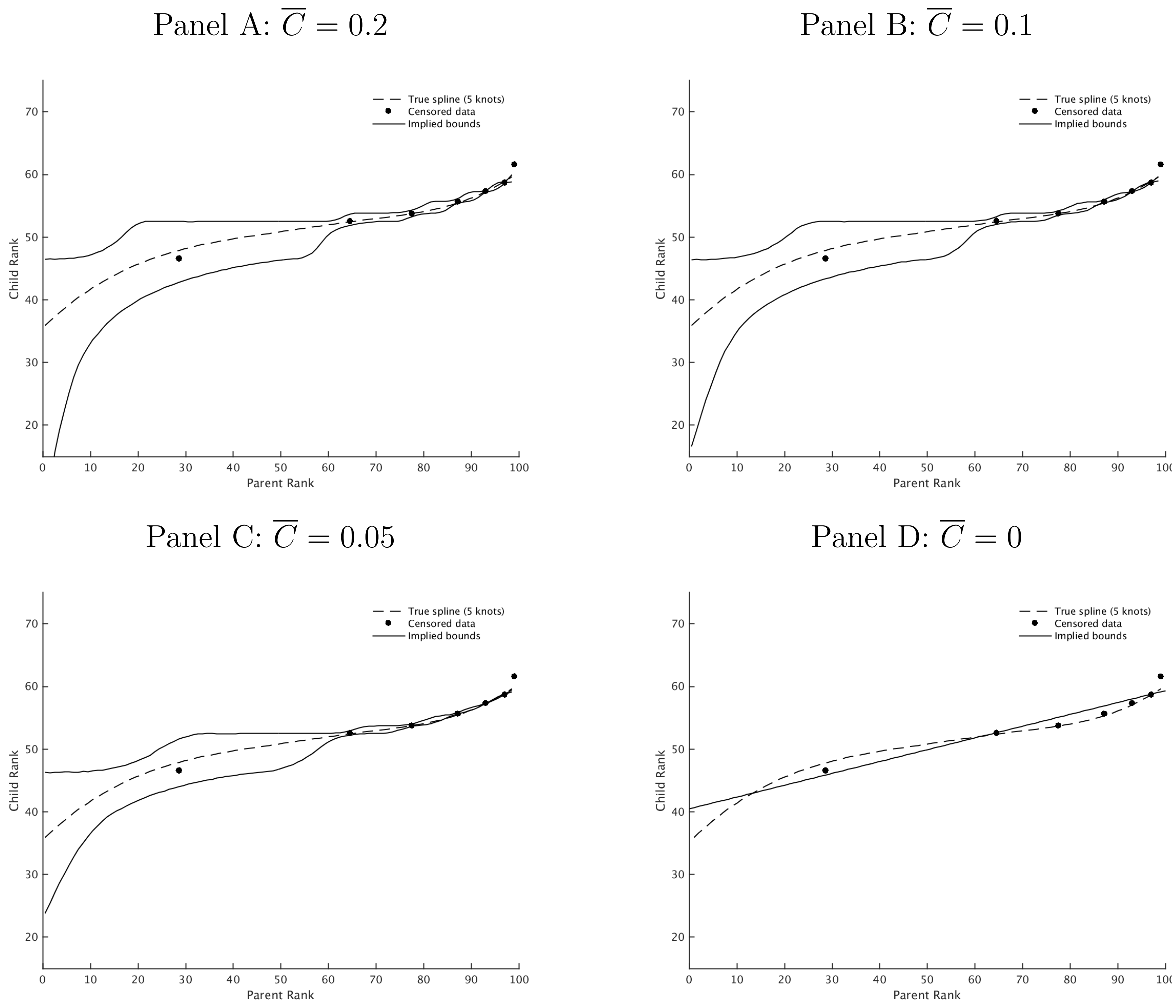
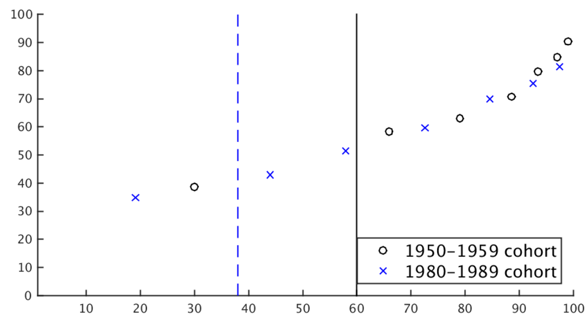



Figure 6 shows results from a simulation using parent-child matched income rank data from Denmark. We simulated interval censoring, so that the only observable data were the points in the graphs, indicating the mean of parent and child rank in each parent bin. We then calculated bounds under four different curvature constraints, indicated in the graph titles. The solid lines show the upper and lower bound of the CEF at each point in the parent distribution, and the dashed line shows the spline fit to the underlying data (described in Figure 5). The data are from Boserup et al. (2014).

**Figure 7**

Changes in Intergenerational Educational Mobility in India
from 1950s to 1980s Birth Cohorts

Panel A: Rank Bin Midpoints




Panel B: CEF Bounds

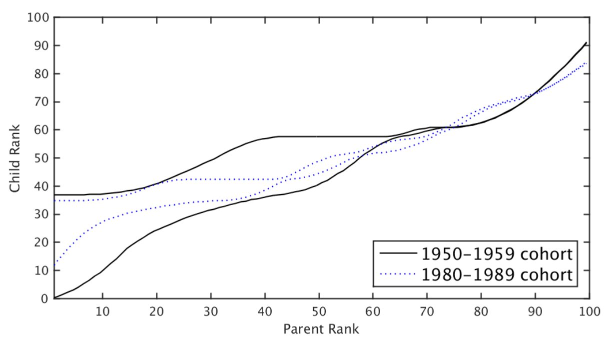
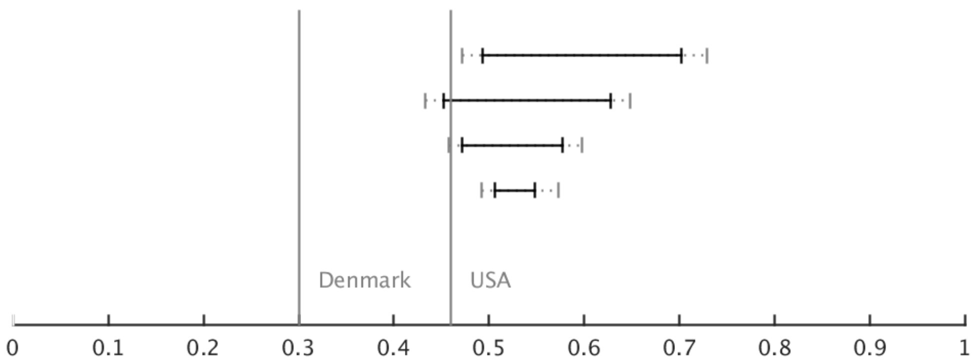



Figure 7 presents the change over time in the rank-rank relationship between Indian fathers and sons born in the 1950s and the 1980s. Panel A presents the raw rank bin means in the data. The vertical lines indicate the size of the lowest parent education rank bin, representing fathers with zero years of education; the solid line shows this value for the 1950s cohort, and the dashed line for the 1980s cohort. Panels B presents the bounds on the child rank conditional expectation function at each parent rank, under our preferred curvature constraint $\overline{C} = 0.10$.

**Figure 8**

Mobility Bounds for 1950s to 1980s Birth Cohorts

Panel A: Rank-Rank Gradient




Panel B: Absolute and Interval Mobility: $p_{25}$ and $\mu_0^{50}$

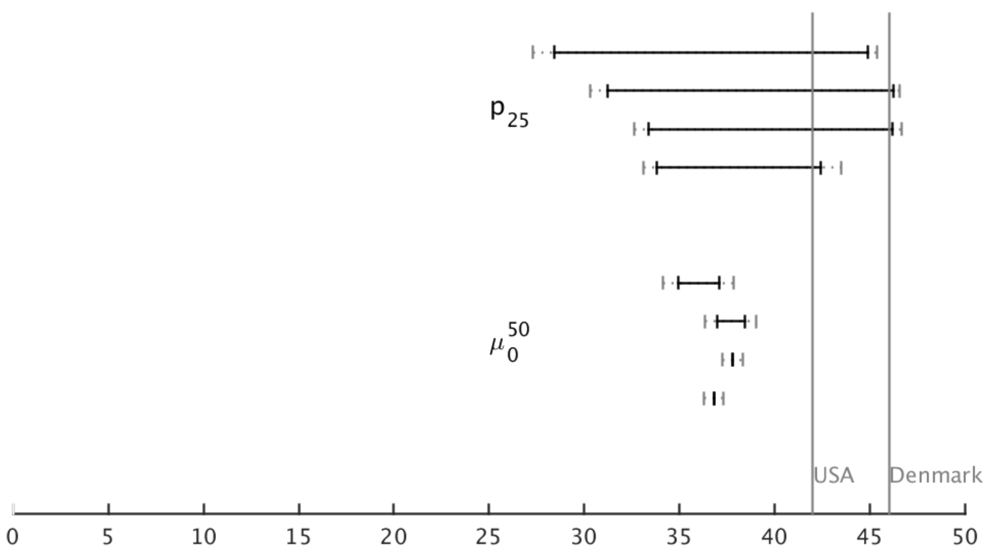
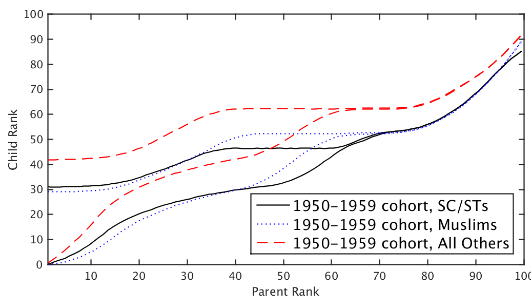



Figure 8 shows bounds on three mobility statistics, estimated on four decades of matched Indian father-son pairs. The solid lines show the estimated bounds on each statistic and the gray dashed lines show the 95% bootstrap confidence sets, based on 1000 bootstrap samples. Each of these statistics was calculated using our preferred curvature constraint of $\overline{C} = 0.10$. For reference, we display the rank-rank education gradient for USA and Denmark (from Hertz (2008)), and $p_{25}$ for USA and Denmark (from Chetty et al. (2014a)). The rank-rank gradient is the slope coefficient from a regression of son education rank on father education rank. $p_{25}$ is absolute upward mobility, which is the expected rank of a son born to a family at the 25th percentile. $\mu_0^{50}$ is upward interval mobility, which is the expected rank of a son born below to a family below the 50th percentile.

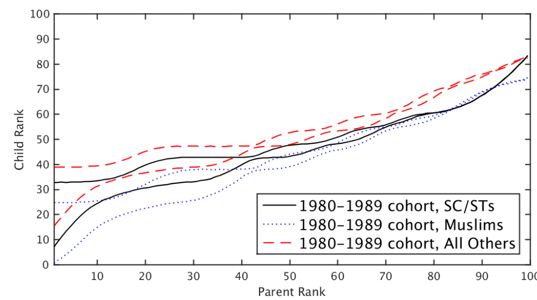
# Figure 9

Bounds on Intergenerational Educational Mobility:
Muslims vs. SC/STs vs. All Others

Panel A: Sons Born in 1950s




Panel B: Sons Born in 1980s




Panel C: Absolute and Interval Mobility

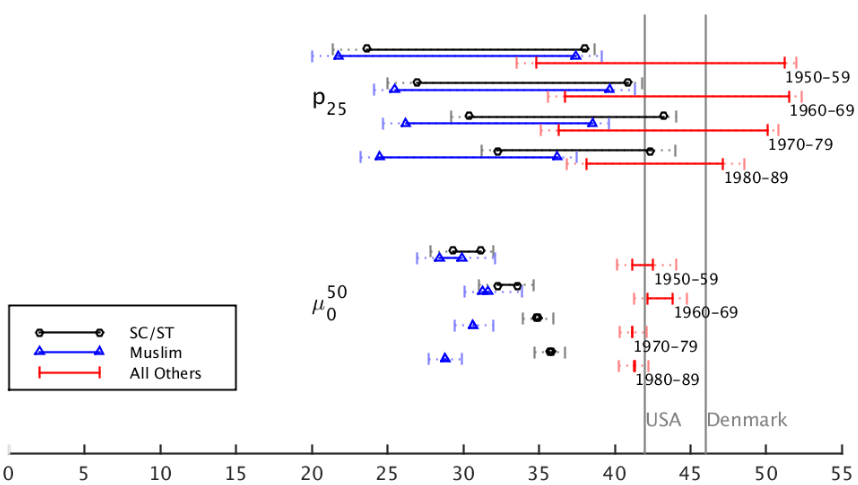



Figure 9 describes the parent-child education rank distribution separately for scheduled caste and scheduled tribes (SC/STs), Muslims and for non-SC/ST non-Muslims, labeled as all others. Panels A and B show bounds on the child rank CEF at every value of the parent rank distribution, under the curvature constraint $\overline{C} = 0.10$. Panel A shows bounds for cohorts born 1950-59 and Panel B shows cohorts born 1980-89. Panel C reports bounds (solid lines) and 95% bootstrap confidence sets (dashed light-colored lines, 1000 bootstrap samples) for two statistics of mobility, calculated using $\overline{C} = 0.10$. Red lines indicate mobility for SC/STs, blue lines for Muslims and black lines for all others. $p_{25}$ is absolute upward mobility, which is the expected rank of a son born to a family at the 25th percentile. $\mu_0^{50}$ is upward interval mobility, which is the expected rank of a son born below to a family below the 50th percentile. We omit the gradient because it measures mobility with respect to the subgroup rather than the national population.

# A Appendix A: Additional Tables and Figures

## Table A1
Transition Matrices

**A: Sons Born 1950-59**

| | Son highest education attained | | | | | | |
|---|---|---|---|---|---|---|---|
| Father ed attained | None (31%) | Some prim. (11%) | Primary (17%) | Middle (13%) | Sec. (13%) | Sr. sec. (6%) | Any higher (8%) |
| None (60%) | 0.47 | 0.12 | 0.17 | 0.11 | 0.09 | 0.03 | 0.03 |
| Some primary (12%) | 0.10 | 0.18 | 0.22 | 0.19 | 0.16 | 0.09 | 0.06 |
| Primary (13%) | 0.07 | 0.08 | 0.31 | 0.16 | 0.19 | 0.08 | 0.10 |
| Middle (6%) | 0.06 | 0.05 | 0.09 | 0.30 | 0.17 | 0.14 | 0.18 |
| Secondary (5%) | 0.03 | 0.02 | 0.04 | 0.12 | 0.37 | 0.11 | 0.30 |
| Sr. secondary (2%) | 0.02 | 0.00 | 0.03 | 0.11 | 0.11 | 0.35 | 0.38 |
| Any higher ed (2%) | 0.01 | 0.01 | 0.01 | 0.03 | 0.08 | 0.13 | 0.72 |

**B: Sons Born 1960-69**

| | Son highest education attained | | | | | | |
|---|---|---|---|---|---|---|---|
| Father ed attained | None (27%) | Some prim. (10%) | Primary (16%) | Middle (16%) | Sec. (14%) | Sr. sec. (7%) | Any higher (10%) |
| None (57%) | 0.41 | 0.12 | 0.16 | 0.14 | 0.09 | 0.04 | 0.04 |
| Some primary (13%) | 0.12 | 0.17 | 0.18 | 0.22 | 0.15 | 0.08 | 0.08 |
| Primary (14%) | 0.09 | 0.05 | 0.26 | 0.18 | 0.20 | 0.09 | 0.13 |
| Middle (6%) | 0.06 | 0.04 | 0.09 | 0.29 | 0.21 | 0.13 | 0.19 |
| Secondary (6%) | 0.03 | 0.02 | 0.08 | 0.12 | 0.35 | 0.16 | 0.25 |
| Sr. secondary (2%) | 0.02 | 0.02 | 0.03 | 0.07 | 0.19 | 0.25 | 0.41 |
| Any higher ed (2%) | 0.01 | 0.01 | 0.02 | 0.03 | 0.09 | 0.11 | 0.73 |

**C: Sons Born 1970-79**

| | Son highest education attained | | | | | | |
|---|---|---|---|---|---|---|---|
| Father ed attained | None (20%) | Some prim. (8%) | Primary (17%) | Middle (18%) | Sec. (16%) | Sr. sec. (10%) | Any higher (12%) |
| None (50%) | 0.33 | 0.10 | 0.19 | 0.17 | 0.12 | 0.05 | 0.04 |
| Some primary (11%) | 0.11 | 0.16 | 0.20 | 0.22 | 0.15 | 0.08 | 0.08 |
| Primary (15%) | 0.08 | 0.06 | 0.24 | 0.23 | 0.18 | 0.11 | 0.11 |
| Middle (8%) | 0.05 | 0.03 | 0.09 | 0.29 | 0.21 | 0.17 | 0.16 |
| Secondary (9%) | 0.03 | 0.02 | 0.06 | 0.12 | 0.31 | 0.19 | 0.27 |
| Sr. secondary (3%) | 0.01 | 0.01 | 0.02 | 0.08 | 0.17 | 0.29 | 0.42 |
| Any higher ed (4%) | 0.00 | 0.00 | 0.02 | 0.05 | 0.10 | 0.17 | 0.66 |

**D: Sons Born 1980-89**

| | Son highest education attained | | | | | | |
|---|---|---|---|---|---|---|---|
| Father ed attained | None (12%) | Some prim. (7%) | Primary (16%) | Middle (20%) | Sec. (16%) | Sr. sec. (12%) | Any higher (17%) |
| None (38%) | 0.26 | 0.10 | 0.21 | 0.20 | 0.12 | 0.06 | 0.05 |
| Some primary (11%) | 0.08 | 0.17 | 0.19 | 0.24 | 0.15 | 0.09 | 0.08 |
| Primary (17%) | 0.05 | 0.04 | 0.22 | 0.23 | 0.20 | 0.13 | 0.13 |
| Middle (12%) | 0.03 | 0.02 | 0.10 | 0.28 | 0.20 | 0.17 | 0.20 |
| Secondary (11%) | 0.02 | 0.01 | 0.05 | 0.13 | 0.23 | 0.24 | 0.32 |
| Sr. secondary (5%) | 0.02 | 0.01 | 0.04 | 0.09 | 0.15 | 0.24 | 0.46 |
| Any higher ed (5%) | 0.01 | 0.01 | 0.02 | 0.05 | 0.10 | 0.16 | 0.65 |

Table A1 shows transition matrices by decadal birth cohort for Indian fathers and sons in the study.

**Figure A1**

Educational attainment by birth cohort and caste

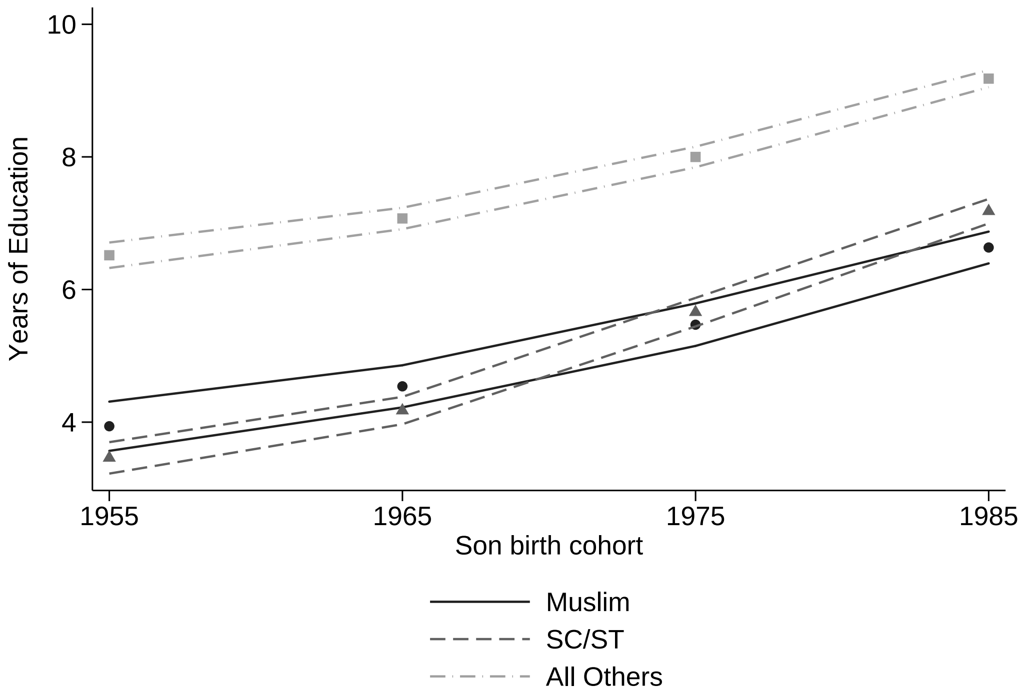
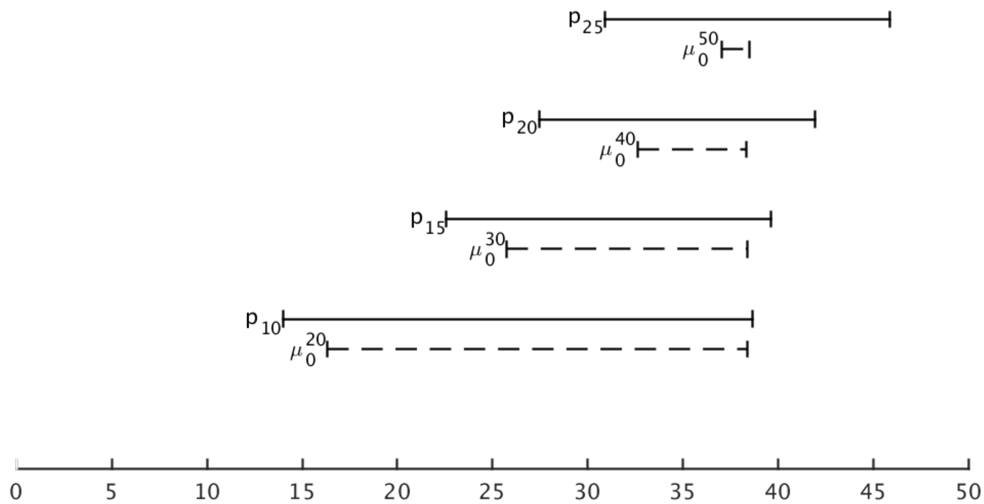



Figure A1 presents average years of education by birth cohort, stratified by members of scheduled castes and tribes, Muslims, and the non-SC/ST non-Muslim population, for all men in India. Data are from the 2011-12 India Human Development Survey. We display birth cohorts from the 1950s, 1960s, 1970s and 1980s. Average outcomes are displayed on the $Y$ axis at the middle of each birth cohort on the $X$ axis. Lines plot 95% confidence intervals.

**Figure A2**

Absolute Mobility and Interval Mobility
at Different Parent Ranks




The graph shows estimates of absolute mobility and interval mobility for Indian sons born 1960-69. Absolute mobility at percentile $p$ is the expected rank of a child born to a family at the $p^{th}$ percentile. Interval mobility below percentile $m$ is the expected rank of a child born to a family *below* the $m^{th}$ percentile. The graph shows estimated bounds on these parameters using the preferred curvature constraint $\overline{C} = 0.10$.

**Figure A3**

Analytical and Numerical Bounds on the Child Rank CEF

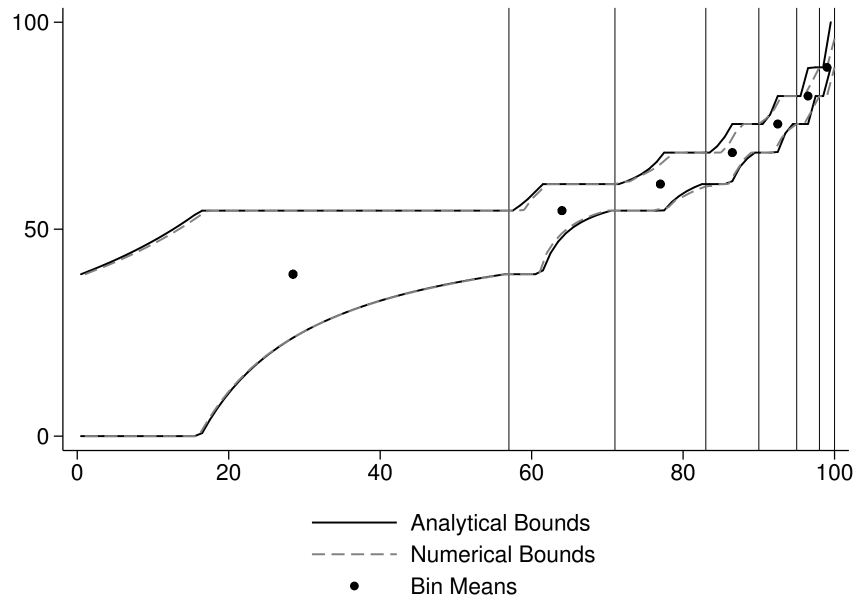
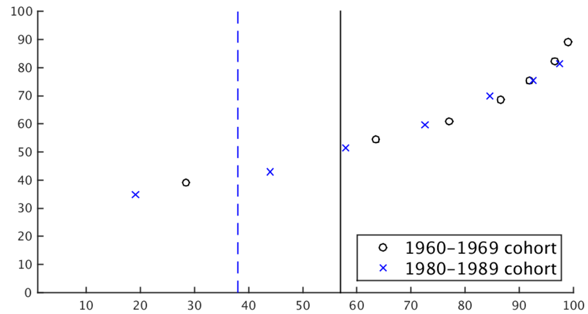



Figure A3 shows bounds on the conditional expectation of child rank given parent rank. The vertical lines show the bin boundaries and the points show the mean value of child rank in each bin. The points are centered at the bin medians. The solid lines show analytical bounds when the distribution of the $X$ variable is uniform. The dashed lines show the numerical bounds calculated with unconstrained curvature, i.e. $\overline{C} = \infty$.

**Figure A4**

Changes in Intergenerational Educational Mobility in India
from 1960s to 1980s Birth Cohorts

Panel A: Rank Bin Medians




Panel B: CEF Bounds

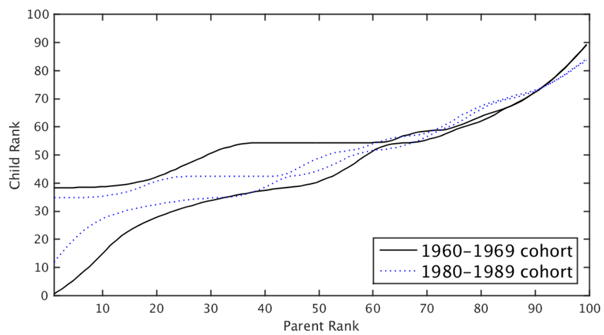
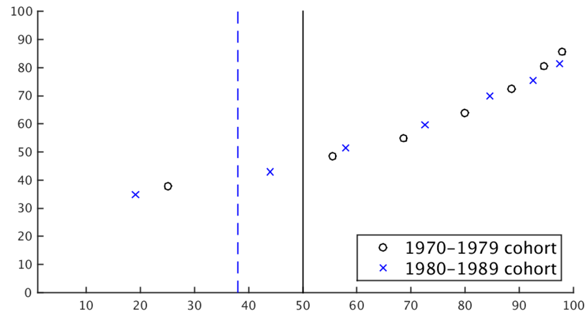



Figure A4 presents the change over time in the rank-rank relationship between Indian fathers and sons born in the 1960s and the 1980s. Panel A presents the rank bin medians in the data. The vertical lines indicate the size of the lowest parent education rank bin, representing fathers with zero years of education; the solid line shows this value for the 1960s cohort, and the dashed line for the 1980s cohort. Panel B presents the bounds on the child rank conditional expectation function at each parent rank, under our preferred curvature constraint $\overline{C} = 0.10$.

**Figure A5**

Changes in Intergenerational Educational Mobility in India from 1970s to 1980s Birth Cohorts

Panel A: Rank Bin Medians




Panel B: CEF Bounds

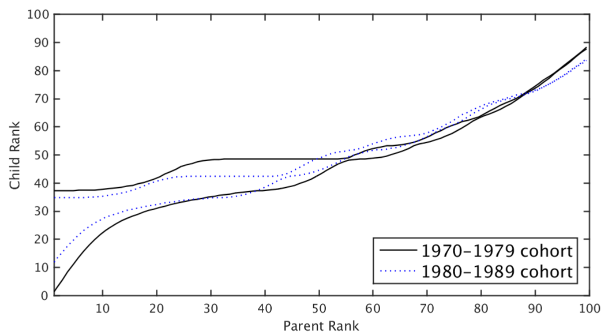
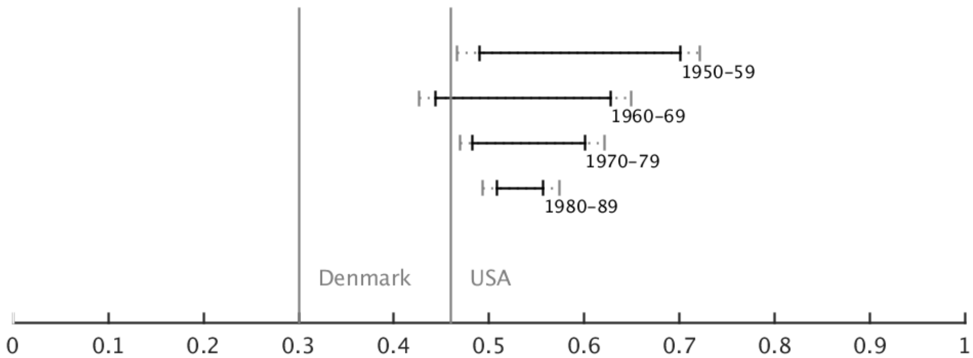



Figure A5 presents the change over time in the rank-rank relationship between Indian fathers and sons born in the 1970s and the 1980s. Panel A presents the rank bin medians in the data. The vertical lines indicate the size of the lowest parent education rank bin, representing fathers with zero years of education; the solid line shows this value for the 1970s cohort, and the dashed line for the 1980s cohort. Panel B presents the bounds on the child rank conditional expectation function at each parent rank, under our preferred curvature constraint $\overline{C} = 0.10$.

## Figure A6

Bounds on Mobility for 1950s to 1980s Birth Cohorts
Estimates from IHDS 2011 Only

Panel A: Rank-Rank Gradient




Panel B: Absolute and Interval Mobility: $p_{25}$ and $\mu_0^{50}$

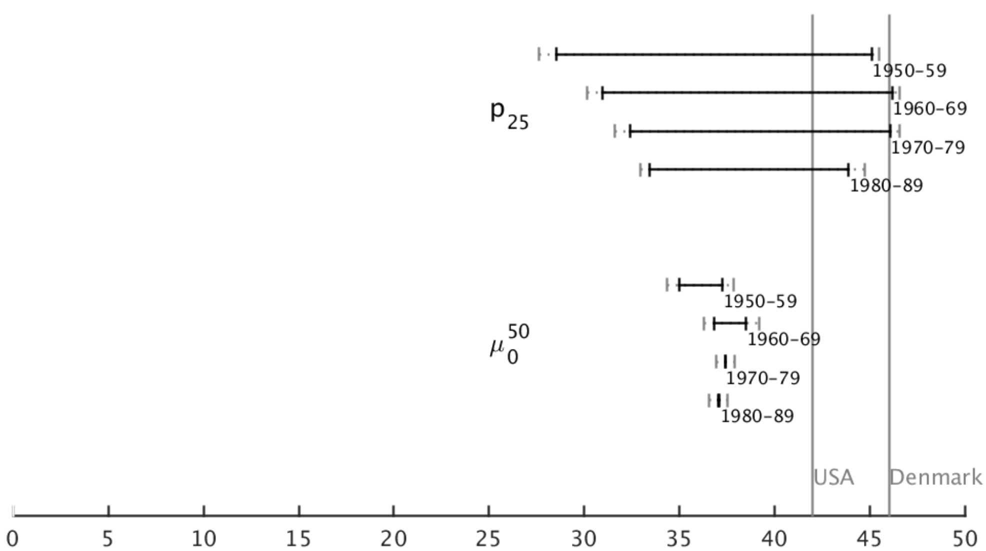


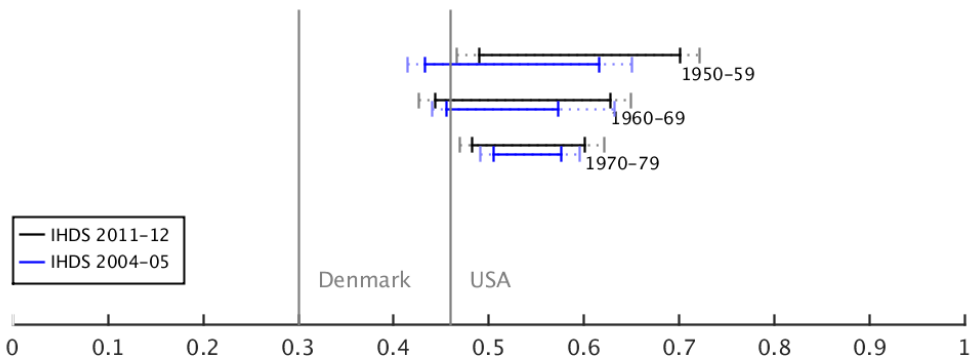

Figure A6 shows bounds on three mobility statistics for matched Indian father-son pairs. Unlike the main estimates in Figure 8, these are based strictly on the IHDS 2011, rather than the joint IHDS-SECC data. Solid lines show the estimated bounds on each statistic and dashed gray lines show the 95% bootstrap confidence intervals, based on 1000 bootstrap samples. Each of these statistics was calculated using the preferred curvature constraint of $\overline{C} = 0.10$. For reference, we display the rank-rank education gradient for USA and Denmark (from Hertz (2008)), and $p_{25}$ for USA and Denmark (from Chetty et al. (2014a)). The rank-rank gradient is the slope coefficient from a regression of son education rank on father education rank. $p_{25}$ is absolute upward mobility, which is the expected rank of a son born to a family at the 25th percentile. $\mu_0^{50}$ is upward interval mobility, which is the expected rank of a son born below to a family below the 50th percentile.

**Figure A7**

Intergenerational Mobility from 1950s to 1980s Birth Cohorts

IHDS 2011 vs. IHDS 2005

Panel A: Rank-Rank Gradient




Panel B: Absolute and Interval Mobility: $p_{25}$ and $\mu_0^{50}$

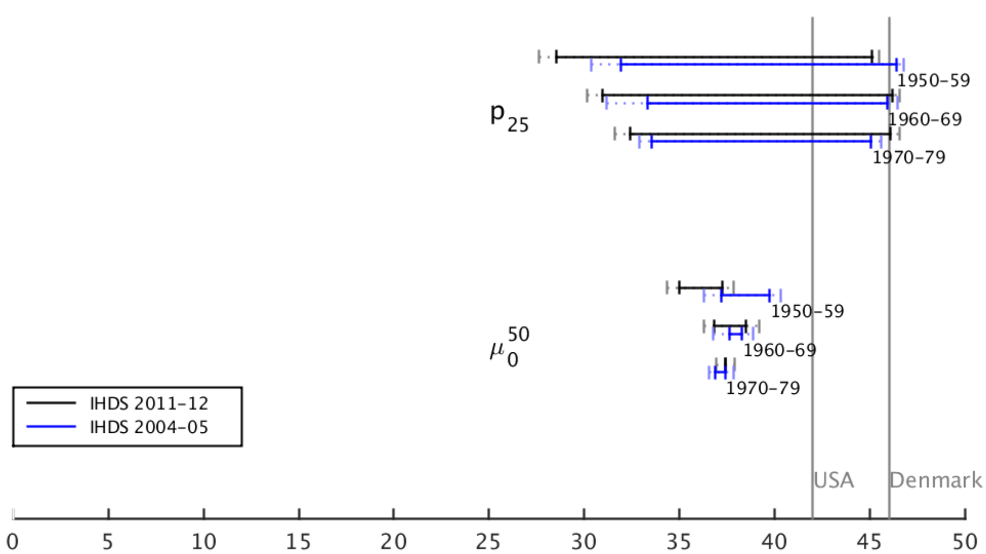
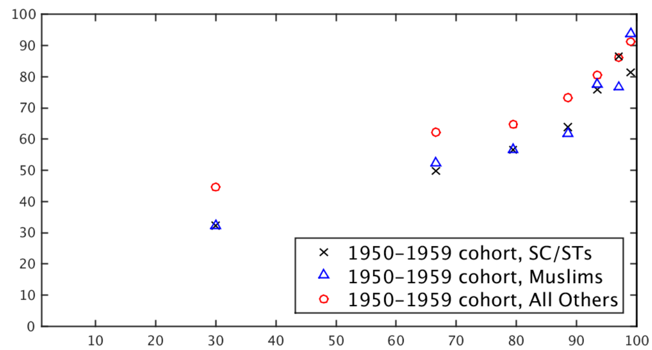



Figure A6 shows bounds on three mobility statistics for matched Indian father-son pairs. For each decadal birth cohort, the two lines show estimates calculated from the IHDS 2005 and the IHDS 2011. Since the birth cohorts are held constant, these reflect education levels of the same cohorts, measured at different points in time. Solid lines show the estimated bounds on each statistic and dashed gray lines show the 95% bootstrap confidence intervals, based on 100 bootstrap samples. Each of these statistics was calculated using our preferred curvature constraint of $\overline{C} = 0.10$. For reference, we display the rank-rank education gradient for USA and Denmark (from Hertz (2008)), and $p_{25}$ for USA and Denmark (from Chetty et al. (2014a)). The rank-rank gradient is the slope coefficient from a regression of son education rank on father education rank. $p_{25}$ is absolute upward mobility, which is the expected rank of a son born to a family at the 25th percentile. $\mu_0^{50}$ is upward interval mobility, which is the expected rank of a son born below to a family below the 50th percentile.

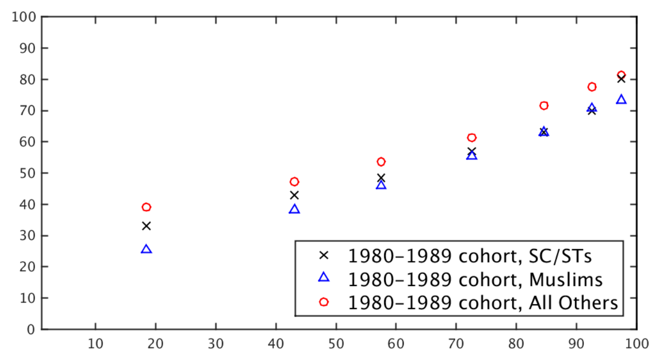
## Figure A8

Intergenerational Educational Mobility in India:
Parent-Child Rank Bin Medians for SC/STs, Muslims and All Others

Panel A: Sons Born in 1950s




Panel B: Sons Born in 1980s




Panel A shows the mean rank bin for sons born in the 1950s, conditioning on the parent rank bin. Parent rank bins are plotted at the midpoint of the bin. Panel B shows the same result for sons born in the 1980s. Results are shown separately for scheduled caste and scheduled tribes (SC/STs), Muslims and for non-SC/ST non-Muslims.

## B Appendix B: Proofs

**Proof of Proposition 1.** Define the conditional expectation function of child rank given parent rank as $Y(i) = E(y|i)$. Let the function $Y(i)$ be defined on $i \in [0, 100]$, and assume $Y(i)$ is integrable. We want to bound $E(y|i)$ when $i$ is known to lie in the interval $[i_k, i_{k+1}]$; there are $K$ such intervals. Define the expected child outcome in bin $k$ as

$$r_k = \frac{1}{i_{k+1} - i_k} \int_{i_k}^{i_{k+1}} Y(i) di.$$

Define $r_0 = 0$ and $r_{K+1} = 100$.

Restate the following assumptions from Manski and Tamer (2002):

$$P(i \in [i_k, i_{k+1}]) = 1. \qquad \text{(Assumption I)}$$

$$E(y|i) \text{ must be weakly increasing in } i. \qquad \text{(Assumption M)}$$

$$E(y|i \text{ is interval censored }) = E(y|i). \qquad \text{(Assumption MI)}$$

From Manski and Tamer (2002), we have:

$$r_{k-1} \le E(y|i) \le r_{k+1} \qquad \text{(Manski-Tamer bounds)}$$

Note also that

$$i \sim U(0, 100) \qquad \text{(Assumption U)}$$

**Proposition 1.** *Let $i$ be in bin $k$. Under assumptions IMMI (Manski and Tamer, 2002) and U, the following bounds on $E(y|i)$ are sharp:*

$$\begin{cases} r_{k-1} \le E(y|i) \le \frac{1}{i_{k+1}-i} \left( (i_{k+1} - i_k) \, r_k - (i - i_k) \, r_{k-1} \right), & i < k^* \\ \frac{1}{i-i_k} \left( (i_{k+1} - i_k) \, r_k - (i_{k+1} - i) \, r_{k+1} \right) \le E(y|i) \le r_{k+1}, & i \ge k^* \end{cases}$$

*where*

$$k^* = \frac{i_{k+1} r_{k+1} - (i_{k+1} - i_k) \, r_k - i_k r_{k-1}}{r_{k+1} - r_{k-1}}.$$

The intuition behind the proof is as follows. First, find the function $z$ which meets the bin mean and is defined as $r_{k-1}$ up to some point $j$. Because $z$ is a valid CEF, the lower bound on $E(y|i)$ is no larger than $z$ up to $j$; we then show that $j$ is precisely $k^*$ from the statement. For points $i > k^*$, we show that the CEF which minimizes the value at point $i$ must be a horizontal line up to $i$ and a horizontal line at $r_{k+1}$ for points larger than $i$. But there is only one such CEF, given that the CEF must also meet the bin mean, and we can solve analytically for the minimum value the CEF can attain at point $i$. We focus on lower bounds for brevity, but the proof for upper bounds follows a symmetric structure.

*Part 1: Find $k^*$.* First define $F_k$ as the set of weakly increasing CEFs which meet the bin mean. Put otherwise, let $F_k$ be the set of $f : [i_k, i_{k+1}] \to \mathbb{R}$ satisfying

$$r_k = \frac{1}{i_{k+1} - i_k} \int_{i_k}^{i_{k+1}} f(i) di.$$

Now choose $z \in F_k$ such that

$$z(i) = \begin{cases} r_{k-1}, & i_k \leq i < j \\ r_{k+1}, & j \leq i \leq i_{k+1}. \end{cases}$$

Note that $z$ and $j$ both exist and are unique (it suffices to show that just $j$ exists and is unique, as then $z$ must be also). We can solve for $j$ by noting that $z$ lies in $F_k$, so it must meet the bin mean. Hence, by evaluating the integrals, $j$ must satisfy:

$$\begin{aligned} r_k &= \frac{1}{i_{k+1} - i_k} \int_{i_k}^{i_{k+1}} z(i) di \\ &= \frac{1}{i_{k+1} - i_k} \left( \int_{i_k}^{j} r_{k-1} di + \int_{j}^{i_{k+1}} r_{k+1} di \right) \\ &= \frac{1}{i_{k+1} - i_k} \left( (j - i_k) r_{k-1} + (i_{k+1} - j) r_{k+1} \right). \end{aligned}$$

Note that these expressions invoke assumption U, as the integration of $z(i)$ does not require any adjustment for the density on the $i$ axis. For a more general proof with an arbitrary distribution of $i$, see section B.A.

With some algebraic manipulations, we obtain that $j = k^*$.

*Part 2: Prove the bounds.* In the next step, we show that $k^*$ is the smallest point at which no $f \in F_k$ can be $r_{k-1}$, which means that there must be some larger lower bound on $E(y|i)$ for $i \geq k^*$. In other words, we prove that

$$k^* = \sup \Big\{ i | \text{ there exists } f \in F_k \text{ such that } f(i) = r_{k-1}. \Big\}.$$

We must show that $k^*$ is an upper bound and that it is the least upper bound.

First, $k^*$ is an upper bound. Suppose that there exists $j' > k^*$ such that for some $g \in F_k$, $g(j') = r_{k-1}$. Observe that by monotonicity and the bounds from Manski and Tamer (2002), $g(i) = r_{k+1}$ for $i \leq j'$; in other words, if $g(j')$ is the mean of the mean of the prior bin, it can be no lower or higher than the mean of the prior bin up to point $j'$. But since $j' > j$, this means that

$$\int_{i_k}^{j'} g(i) di < \int_{i_k}^{j'} z(i) di,$$

since $z(i) > g(i)$ for all $h \in (j, j')$. But recall that both $z$ and $g$ lie in $F_k$ and must therefore meet the bin mean; i.e.,

$$\int_{i_k}^{i_{k+1}} g(i) di = \int_{i_k}^{i_{k+1}} z(i) di.$$

But then

$$\int_{j'}^{i_{k+1}} g(i) di > \int_{j'}^{i_{k+1}} z(i) di.$$

That is impossible by the bounds from Manski and Tamer (2002), since $g(i)$ cannot exceed $r_{k+1}$, which is precisely the value of $z(i)$ for $i \geq j$.

Second, $j$ is the least upper bound. Fix $j' < j$. From the definition of $z$, we have shown that for some $h \in (j', j)$, $z(h) = r_{k-1}$ (and $z \in F_k$). So any point $j'$ less than $j$ would not be a lower bound on the set — there is a point $h$ larger than $j'$ such that $z(h) = r_{k-1}$.

Hence, for all $i < k^*$, there exists a function $f \in F_k$ such that $f(i) = r_{k-1}$; the lower bound on $E(y|i)$ for $i < k^*$ is no greater than $r_{k-1}$. By choosing $z'$ with

$$z'(i) = \begin{cases} r_{k-1}, & i_k \leq i \leq j \\ r_{k+1}, & j < i \leq i_{k+1}, \end{cases}$$

it is also clear that at $k^*$, the lower bound is no larger than $r_{k-1}$ (and this holds in the proposition itself, substituting in $k^*$ into the lower bound in the second equation).

Now, fix $i' \in (k^*, i_{k+1}]$. Since $k^*$ is the supremum, there is no function $f \in F_k$ such that $f(i') = r_{k-1}$. Thus for $i' > k^*$, we seek a sharp lower bound larger than $r_{k-1}$. Write this lower bound as

$$Y_{i'}^{min} = \min \Big\{ f(i') \text{ for all } f \in F_k \Big\},$$

where $Y_{i'}^{min}$ is the smallest value attained by any function $f \in F_k$ at the point $i'$.

We find this $Y_{i'}^{min}$ by choosing the function which maximizes every point after $i'$, by attaining the value of the subsequent bin. The function which minimizes $f(i')$ must be a horizontal line up to this point.

Pick $\tilde{z} \in F_k$ such that

$$\tilde{z}(i) = \begin{cases} \underline{Y}, & i_k \leq i' \\ r_{k+1}, & i' < i_{k+1} \leq i_{k+1} \end{cases}.$$

By integrating $\tilde{z}(i)$, we claim that $\underline{Y}$ satisfies the following:

$$\frac{1}{i_{k+1} - i_k} \left( (i' - i_k)\, \underline{Y} + (i_{k+1} - i')\, r_{k+1} \right) = r_k.$$

As a result, $\underline{Y}$ from this expression exists and is unique, because we can solve the equation. Note that this integration step also requires that the distribution of $i$ be uniform, and we generalize this argument in B.A.

By similar reasoning as above, there is no $Y' < \underline{Y}$ such that there exists $g \in F_k$ with $g(i') = Y'$. Otherwise there must be some point $i > i'$ such that $g(i') > r_{k+1}$ in order that $g$ matches the bin means and lies in $F_k$; the expression for $\underline{Y}$ above maximizes every point after $i'$, leaving no additional room to further depress $\underline{Y}$.

Formally, suppose there exists $g \in F_k$ such that $g(i') = Y' < \underline{Y}$. Then $g(i') < \tilde{z}(i')$ for all $i < i'$, since $g$ is monotonic. As a result,

$$\int_{i_k}^{i'} \tilde{z}(i) di > \int_{i_k}^{i'} g(i) di.$$

But recall that

$$\int_{i_k}^{i_{k+1}} g(i) di = \int_{i_k}^{i_{k+1}} \tilde{z}(i) di,$$

so

$$\int_{i'}^{i_{k+1}} g(i)di > \int_{i'}^{i_{k+1}} \tilde{z}(i)di.$$

This is impossible, since $\tilde{z}(i) = r_{k+1}$ for all $i > i'$, and by Manski and Tamer (2002), $g(i) \leq r_{k+1}$ for all $g \in F_k$. Hence there is no such $g \in F_k$, and therefore $\underline{Y}$ is smallest possible value at $i'$, i.e. $\underline{Y} = Y_{i'}^{min}$.

By algebraic manipulations, the expression for $\underline{Y} = Y_i^{min}$ reduces to

$$Y_i^{min} = \frac{(i_{k+1} - i_k)\, r_k - (i_{k+1} - i) r_{k+1}}{i - i_k},\ i \geq k^*.$$

The proof for the upper bounds uses the same structure as the proof of the lower bounds.

Finally, the body of this proof gives sharpness of the bounds. For we have introduced a CEF $f \in F_k$ that obtains the value of the upper and lower bound for any point $i \in [i_k, i_{k+1}]$. For any value $y$ within the bounds, one can generate a CEF $f \in F_k$ such that $f(i) = y$. $\square$

Appendix Figure B1 gives a visual depiction of the analytical CEFs that generate the bounds on $p_{20}$ and $p_{25}$ respectively for the cohort born in the 1960s in our India example. Both $i = 20$ and $i = 25$ lie in bin 1. For each of these, the upper and lower bounds are generated by the same CEFs. The CEF that maximizes at $p_{20}$ is a horizontal line at $r_{k+1} = 54.48$ for all values after $i = 20$, because $i = 20$ lies after $k^* = 16.18$. The CEF that minimizes $p_{20}$ must be the smallest horizontal line up to $i = 20$ such that the average value in the first bin is 39.13. The same logic holds for the CEF that minimizes/maximizes $i = 25$, but the lower bound is higher because if any of the points below 25 were made any lower, it would be impossible for the average value in the first bin to match the data.

**Proof of Proposition 2**. Define

$$\mu_a^b = \frac{1}{b-a} \int_a^b E(y|i)di.$$

Let $Y_i^{min}$ and $Y_i^{max}$ be the lower and upper bounds respectively on $E(y|i)$ given by proposition 1. We seek to bound $\mu_a^b$ when $i$ is observed only in discrete intervals.

**Proposition 2.** *Let $b \in [i_k, i_{k+1}]$ and $a \in [i_h, i_{h+1}]$. Let assumptions IMMI (Manski and Tamer, 2002) and U hold. Then the following bounds are sharp:*

$$\begin{cases} Y_b^{min} \leq \mu_a^b \leq Y_a^{max}, & h = k \\ \frac{r_h(i_k - a) + Y_b^{min}(b - i_k)}{b-a} \leq \mu_a^b \leq \frac{Y_a^{max}(i_k - a) + r_k(b - i_k)}{b-a}, & h+1 = k \\ \frac{r_h(i_{h+1} - a) + \sum_{\lambda=h+1}^{k-1} r_\lambda(i_{\lambda+1} - i_\lambda) + Y_b^{min}(b - i_k)}{b-a} \leq \mu_a^b \leq \frac{Y_a^{max}(i_{h+1} - a) + \sum_{\lambda=h+1}^{k-1} r_\lambda(i_{\lambda+1} - i_\lambda) + r_k(b - i_k)}{b-a}, & h+1 < k. \end{cases}$$

The order of the proof is as follows. If $a$ and $b$ lie in the same bin, then $\mu_a^b$ is maximized only if the CEF is minimized prior to $a$. As in the proof of proposition 1, that occurs when the CEF is a horizontal line at $Y_i^{min}$ up to $a$, and a horizontal line $Y_i^{max}$ at and after $a$. If

$a$ and $b$ lie in separate bins, the value of the integral in bins that are contained between $a$ and $b$ is determined by the observed bin means. The portions of the integral that are not determined are maximized by a similar logic, since they both lie within bins. We prove the bounds for maximizing $\mu_a^b$, but the proof is symmetric for minimizing $\mu_a^b$.

*Part 1: Prove the bounds if $a$ and $b$ lie in the same bin.* We seek to maximize $\mu_a^b$ when $a, b \in [i_k, i_{k+1}]$. This requires finding a candidate CEF $f \in F_k$ which maximizes $\int_a^b f(i)di$. Observe that the function $f(i)$ defined as

$$f(i) = \begin{cases} Y_a^{min}, & i_k \leq i < a \\ Y_a^{max}, & a \leq i \leq i_{k+1} \end{cases}$$

has the property that $f \in F_k$. For if $a \geq k^*$, $f = \tilde{z}$ from the second part of the proof of proposition 1. If $a < k^*$, the CEF in $F_k$ which yields $Y_a^{max}$ is precisely $f$ (by a similar argument which delivers the upper bounds in proposition 1).

This CEF maximizes $\mu_a^b$, because there is no $g \in F_k$ such that

$$\frac{1}{b-a}\int_a^b g(i)di > \frac{1}{b-a}\int_a^b f(i)di.$$

Note that for any $g \in F_k$, $\frac{1}{i_{k+1}-i_k}\int_{i_k}^{i_{k+1}} g(i)di = \frac{1}{i_{k+1}-i_k}\int_{i_k}^{i_{k+1}} f(i)di = r_k$. Hence in order that $\int_a^b g(i)di > \int_a^b f(i)di$, there are two options. The first option is that

$$\int_{i_k}^{a} g(i)di < \int_{i_k}^{a} f(i)di.$$

That is impossible, since there is no room to depress $g$ given the value of $f$ after $a$. If $a < k^*$, then it is clear that there is no $g$ giving a larger $\mu_a^b$, since $r_{k-1} \leq g(i)$ for $i_{k-1} \leq i \leq a$, so $g$ is bounded below by $f$. If $a \geq k^*$, then $f(i) = r_{k+1}$ for all $a \leq i \leq i_{k+1}$. That would leave no room to depress $g$ further; if $\int_{i_k}^a g(i)di < \int_{i_k}^a f(i)di$, then $\int_a^{i_{k+1}} g(i)di > \int_a^{i_{k+1}} f(i)di$, which cannot be the case if $f = r_{k+1}$, by the bounds given in Manski and Tamer (2002).

The second option is that

$$\int_b^{i_k} g(i)di < \int_b^{i_k} f(i)di.$$

This is impossible due to monotonicity. For if $\int_a^b g(i)di > \int_a^b f(i)di$, then there must be some point $i' \in [a, b)$ such that $g(i') > f(i')$. By monotonicity, $g(i) > f(i)$ for all $i \in [i', i_{k+1}]$ since $f(i) = Y_a^{max}$ in that interval. As a result,

$$\int_b^{i_k} g(i)di > \int_b^{i_k} f(i)di,$$

since $b \in (i', i_{k+1})$. (If $b = i_{k+1}$, then only the first option would allow $g$ to maximize the desired $\mu_a^b$.)

Therefore, there is no such $g$, and $f$ indeed maximizes the desired integral. Integrating $f$ from $a$ to $b$, we obtain that the upper bound on $\mu_a^b$ is $\frac{1}{b-a}\int_a^b Y_a^{max}di = Y_a^{max}$. Note that there may be many functions which maximize the integral; we only needed to show that $f$ is one of them.

To prove the lower bound, use an analogous argument.

*Part 2: Prove the bounds if $a$ and $b$ do not lie in the same bin.* We now generalize the set up and permit $a, b \in [0, 100]$. Let $F$ be the set of weakly increasing functions such that $\frac{1}{i_{k+1}-i_k}\int_{i_k}^{i_{k+1}} f(i)di = r_k$ for all $k \leq K$. In other words, $F$ is the set of functions which meet the means of every bin. Now observe that for all $f \in F$,

$$\begin{aligned} \mu_a^b &= \frac{1}{b-a}\int_a^b f(i)di \\ &= \frac{1}{b-a}\left(\int_a^{i_{h+1}} f(i)di + \int_{i_{h+1}}^{i_k} f(i)di + \int_{i_k}^{b} f(i)di\right), \end{aligned}$$

by a simple expansion of the integral.

But for all $f \in F$,

$$\int_{i_{h+1}}^{i_k} f(i)di = \sum_{\lambda=h+1}^{k-1} r_\lambda(i_{\lambda+1} - i_\lambda)$$

if $h+1 < k$ and

$$\int_{i_{h+1}}^{i_k} f(i)di = 0$$

if $h+1 = k$. For in bins completely contained inside $[a, b]$, there is no room for any function in $F$ to vary; they all must meet the bin means.

We proceed to prove the upper bound. We split this into two portions: we wish to maximize $\int_a^{i_{h+1}} f(i)di$ and we also wish to maximize $\int_{i_k}^{b} f(i)di$. The values of these objects are not codependent. But observe that the CEFs $f \in F_k$ which yield upper bounds on these integrals are the very same functions which yield upper bounds on $\mu_a^{i_{h+1}}$ and $\mu_{i_k}^b$, since $\mu_s^t = \frac{1}{t-s}\int_s^t f(i)di$ for any $s$ and $t$. Also notice that $a$ and $i_{h+1}$ both lie in bin $h$, while $b$ and $i_k$ both lie in bin $k$, so we can make us of the first portion of this proof.

In part 1, we showed that the function $f \in F$, $f : [i_h, i_{h+1}] \to \mathbb{R}$, which maximizes $\mu_a^{i_{h+1}}$ is

$$f(i) = \begin{cases} Y_a^{min}, & i_h \leq i < a \\ Y_a^{max}, & a \leq i \leq i_{h+1}. \end{cases}$$

As a result

$$\max_{f\in F}\left\{\int_a^{i_{h+1}} f(i)di\right\} = \int_a^{i_{h+1}} Y_a^{max} di = Y_a^{max}(i_{h+1} - a).$$

Similarly, observe that $i_k$ and $b$ lie in the same bin, so the function $f : [i_k, i_{k+1}] \to \mathbb{R}$, with $f \in F$ which maximizes $\int_{i_k}^b f(i)di$ must be of the form

$$f(i) = \begin{cases} Y_{i_k}^{min}, & i_k \leq i < a \\ Y_{i_k}^{max}, & b \leq i \leq i_{k+1}. \end{cases}$$

With identical logic,

$$\max_{f \in F} \left\{ \int_{i_k}^{b} f(i) di \right\} = \int_{i_k}^{b} Y_{i_k}^{max} di = Y_{i_k}^{max}(b - i_k).$$

And by proposition 1, $i_k \leq k^*$ so $Y_{i_k}^{max} = r_k$. (Note that if $i_k = k^*$, substituting $k^*$ into the second expression of proposition 1 still yields that $Y_{i_k}^{max} = r_k$.)

Now we put all these portions together. First let $h + 1 = k$. Then $\int_{i_{h+1}}^{i_k} f(i)di = 0$, so we have that we maximize $\mu_a^b$ by

$$\frac{1}{b-a} \left( Y_a^{max}(i_{h+1} - a) + r_k(b - i_k) \right).$$

Similarly, if $h + 1 < k$ and there are entire bins completely contained in $[a, b]$, then we maximize $\mu_a^b$ by

$$\frac{1}{b-a} \left( Y_a^{max}(i_{h+1} - a) + \sum_{\lambda=h+1}^{k-1} r_\lambda(i_{\lambda+1} - i_\lambda) + r_k(b - i_k) \right).$$

The lower bound is proved analogously. Sharpness is immediate, since we have shown that the CEF which delivers the endpoints of the bounds lies in $F$. As a result, there is a function delivering any intermediate value for the bounds. □

### B.A Analytical Bounds with an Arbitrary Probability Distribution

**Proposition 3.** *Let $i$ be in bin $k$. Let $p_k(i)$ be the probability density function of $i$ in bin $k$. Under assumptions IMMI (Manski and Tamer, 2002), the following bounds on $E(y|i)$ are sharp:*

$$\begin{cases} r_{k-1} \leq E(y|i) \leq \frac{r_k - r_{k-1} \int_{i_k}^{i} p_k(s)ds}{\int_{i}^{i_{k+1}} p_k(s)ds}, & i < k^* \\ \frac{r_k - r_{k+1} \int_{i}^{i_{k+1}} p_k(s)ds}{\int_{i_k}^{i} p_k(s)ds} \leq E(y|i) \leq r_{k+1}, & i \geq k^* \end{cases}$$

*where $k^*$ satisfies:*

$$r_k = r_{k-1} \int_{i_k}^{k^*} p_k(s)ds + r_{k+1} \int_{k^*}^{i_{k+1}} p_k(s)ds.$$

**Proof of Proposition 3.** The proof follows the same argument as in proposition 1. With an arbitrary distribution, $F_k$ now constitutes the functions $f : [i_k, i_{k+1}] \to \mathbb{R}$ which satisfy:

$$\int_{i_k}^{i_{k+1}} f(s)p_k(s)ds = r_k.$$

As before, choose $z \in F_k$ such that

$$z(i) = \begin{cases} r_{k-1}, & i_k \leq i < j \\ r_{k+1}, & j \leq i \leq i_{k+1}. \end{cases}$$

Because the distribution of $i$ is no longer uniform, in that case, $j$ must satisfy

$$\begin{aligned} r_k &= \int_{i_k}^{i_{k+1}} z(s)p_k(s)ds \\ &= r_{k-1}\int_{i_k}^{j} p_k(s)ds + r_{k+1}\int_{j}^{i_{k+1}} p_k(s)ds. \end{aligned}$$

This implies that $j = k^*$, precisely.

The rest of the arguments follow identically, except we now claim that for $i > k^*$, $\underline{Y} = Y_i^{min}$ satisfies the following:

$$r_k = \int_{i_k}^{i} Y_i^{min} p_k(s)ds + \int_{i}^{i_{k+1}} r_{k+1}p_k(s)ds.$$

By algebraic manipulations, we obtain:

$$Y_i^{min} = \frac{r_k - r_{k+1}\int_i^{i_{k+1}} p_k(s)ds}{\int_{i_k}^{i} p_k(s)ds}$$

and the proof of the lower bounds is complete. As before, the proof for upper bounds follows from identical logic. $\square$

**Figure B1**

Illustration of Proof of Proposition 1

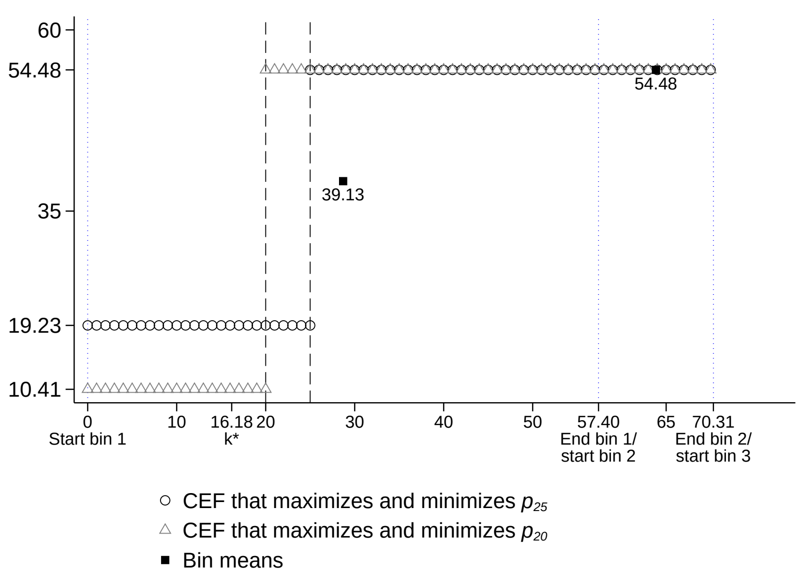



Figure B1 illustrates the CEFs that generate worst-case bounds on $p_{20}$ and $p_{25}$.

## C Appendix C: Mobility Bounds when Child Rank is Interval Censored and Parent Rank is Known

In the main part of the paper, we focus on bounding a function $Y(i) = E(y|i)$ when $Y$ is observed without error, but $i$ is observed with interval censoring. In this section, we modify the setup to allow interval censoring in the observed outcomes $Y$. For simplicity, we focus on a setting where $Y$ is observed with interval censoring but $i$ is observed precisely, and we discuss in the text how the setup could be modified to consider interval censoring in both variables simultaneously.

We begin by setting up a minimization problem that generates bounds on an integrable, weakly monotonically increasing function $Y(i) = E(y|i)$, when $i$ is observed without error, but $Y$ is known only to lie within an interval $[Y_k, Y_{k+1}]$. Notation used is fully described in this section, and does not correspond directly to concepts used in Section IV. Without loss of generality, we take $Y$ to represent child rank and $i$ to represent parent rank in some latent social rank distribution. Given bounds on $E(y|i)$, we can calculate any measure of intergenerational mobility that is a function of the conditional expectation of child rank given parent rank; we therefore continue to focus on bounding the conditional expectation function.

We assume that parents' ranks are precisely observed in $I$ bins, indexed by $i$; we treat the midpoint parent rank in each bin as the true parent rank. Define $h_i$ as the share of parents in each bin; if we observed parent ranks in integer rank bins, $h_i$ would be equal for all values of $i$. Children's ranks are observed in $K$ bins, where bin $k$ is defined by its boundaries $Y_k$ and $Y_{k+1}$.

The joint parent-child rank distribution can be fully described by a set of conditional child CDFs for each parent rank $i$:

$$F_i(r) = P(\text{child rank} \leq r | I = i) \tag{C.1}$$

The function describes the probability that a child born to a parent with rank $i$ attains less than or equal to rank $r$ in the latent outcome distribution. The CEF of child rank given a parent with rank $i$ is given by:

$$E(y|i) = \int_0^{100} r f_i(r) dr, \tag{C.2}$$

where $f_i(r)$ is the PDF corresponding to the CDF in Equation C.1.

Because we do not observe children's ranks exactly, calculating the value of a candidate CEF at some parent rank thus requires an assumption about how the distribution of children within each child education bin depends on parent education. Relative to the case with only parent censoring developed in Section IV, this problem is $I$-dimensional – before bounding the CEF, we need to bound $I$ individual child CDFs corresponding to each parent rank.

The sample analog of $F_i(r)$ is the share of children who attain rank $r$ or lower, conditional on being born to a parent with rank $i$. We call this value $\hat{F}_i(r)$. Because the child data are interval censored, we do not observe $\hat{F}_i(r)$ at every value of $r$ — we observe $\hat{F}_i(r)$ only at the boundaries of the child bins 1 through $K$. However, we observe it precisely at these boundaries.

Consider the following example. In the 1960s birth cohort in India, 26.8% of sons report having completed zero years of education, and an additional 10.1% have started but not completed primary. Therefore, the 26th percentile son has zero years of education, while the

27th percentile son has started but not completed primary education. Among sons of zero-education fathers (who make up 57% of the population), 41% of sons have zero education. Since the 41st percentile son in this group is right at the margin of starting (but not completing) primary school, then under the assumption of a continuous CDF, we know that this son is exactly at the 26.8th percentile of the national latent education rank distribution, which is to say that $\hat{F}_{i_0}(r) = 0.268$, where $i_0$ represents the rank of fathers with zero education. The fact that the data pins down the value of the CDF precisely at these points highlights one dimension under which our bounds under censored child data will be tighter than our bounds under censored parent data. It allows us to estimate with parameters under a mean squared error criterion rather than a integral mean squared error as in Section IV.

The parent-child rank distribution is characterized by a child CDF for each parent rank $i$. We parameterize each of these child CDFs as $F_i(r) = S_i(r, \gamma_i)$, where $r$ is the child rank, $i$ is the parent rank, and $\gamma_i$ is a parameter vector in parameter space G. Similarly let $f_i(r) = s_i(r, \gamma_i)$. In our numerical calculation, we define G as $[0, 100]^{100}$, a vector which gives the value of the child CDF at each of 100 child ranks, but other parameterizations could be used. Note that there are as many vectors $\gamma_i$ as there are parent ranks.

We can obtain the set of feasible values of $\gamma_i$ from the following minimization problem:

$$\Gamma = \text{argmin}_{g \in G} \left[ \sum_{i=1}^{I} \sum_{k=1}^{K} S_i(k, g_i) - \hat{F}_i(k) \right]^2 \tag{C.3}$$

$$\text{such that}$$

$$\frac{s_i(r, g_i)}{1 - S_i(r, g_i)} \text{ is weakly decreasing in } i \qquad \text{(Monotonicity)}$$

$$\sum_{i=1}^{I} h_i S_i(r, g_i) = r \qquad \text{(Budget Constraint)}$$

$$S_i(0, g_i) = 0 \qquad \text{(End Points)}$$

$$S_i(100, g_i) = 1.$$

This minimization problem defines feasible sets of $I$ CDFs that represent the distribution of child outcomes for each parent rank $i$. The monotonicity assumption requires that the hazard function is decreasing in parent rank. This is equivalent to stating that *conditional upon a child being above some point $x$ in the rank distribution*, the probability density must be weakly decreasing in parent rank.[61] The budget constraint requires that the CDF values of each parent group (weighted by parent density) must add up to the population CDF. In other words, $J$% of children must on average attain less than or equal to the $J^{th}$ percentile. The constraints on the end points of the CDF are redundant given the other constraints, but are included to highlight how the end points constrain the set of possible outcomes. For simplicity, we have

[61] This assumption implies that the child rank distribution of a higher ranked parent stochastically dominates the child rank distribution of a lower ranked parent. An example of this assumption in discrete space is to require that among all children with zero education, the children of zero education parents cannot on average be ranked above children of higher education parents. Dardanoni et al. (2012) find that a similar conditional monotonicity holds in almost all mobility tables in 35 countries.

not included a curvature constraint, but as will be evident from the solutions to this problem, such a constraint would be a sensible further restriction on the feasible parameter space.

Once a set of candidate CDFs have been identified, they have a one-to-one correspondence with the child rank CEF given parent rank (described by Equation C.2), and thus with any mobility statistic that is a function of the CEF. These statistics can be numerically bounded as in Section IV.

### C.A Numerical Optimization and Results

This problem is computationally more challenging than the problem of parent rank censoring dealt with in Section IV, because it requires simultaneously optimizing over $I$ distinct 100-valued CDFs. There are $I*100$ parameters, $(I-1)*100$ monotonicity inequality constraints, 100 equality budget constraints, and an additional $I*98$ curvature constraint inequalities if desired.

In principle, a similar problem setup could be used to simultaneously solve for interval censoring in both the parent and child distributions, for example, by identifying a set of feasible CDF values for each of 100 integer parent ranks. This problem has 10,000 parameters and 10,000 constraints, and proved computationally too difficult to resolve. Instead, we limit ourselves in this section to identifying feasible CDFs for children born into the seven parent rank bins that we observe in the data, and assume that the true parent rank is at the midpoint of the parent rank interval that we observe. The numerical optimization therefore generates seven 100-valued functions representing $\gamma_i$ for seven values of $i$.[62]

As an example, we calculate bounds on the parent-child distribution using father and son education from the 1960s Indian birth cohort. The raw data are displayed in the 2nd transition matrix in Table A1, and graphed in Appendix Figure C1. Because the raw data are monotonic, many feasible sets of son CDFs will fit the data with zero MSE. We therefore require that MSE is equal to zero, and set up an optimizer to respectively minimize and maximize the value of some mobility statistic (such as the rank-rank gradient) across all feasible son CDFs.

Appendix Figure C2 shows two CDFs that respectively minimize and maximize the rank-rank gradient.[63] In the parent-child distribution that minimizes mobility (Panel A), children's ranks within each parent-child bin are perfectly ordered by parent rank. For example, among sons who report completing middle school, these CDFs would be generated by the assumption that those who came from low education parents have all just barely completed middle school and thus have the lowest ranks in this subgroup, while those who come from high education parents are very close to attaining the next level of education, and thus have the highest ranks in this subgroup. The step function shape of these CDFs highlights that perfect ordering within child education bins would imply a set of child rank probability density functions that do not appear very natural. If child ranks have a continuous distribution, we would expect the density of child ranks at the bottom ranks of one schooling level to be

[62]We implement the monotonicity assumption by requiring it to hold within each observed bin. This is a conservative implementation; requiring this assumption at every point in the discrete space would be computationally more intensive but would obtain even tighter bounds than what we show here.

[63]The set of CDFs that bound $p_{25}$ and $\mu_{50}$ are highly similar. Minimizing the gradient implies maximizing the CEF of children born to the worst off parents, which effectively maximizes $p_{25}$ and $\mu_{50}$ given the parent ranks we use here.

similar to the density at the top ranks of the lower schooling level. Panel B shows the set of CDFs that maximize mobility given the data. This solution would be generated from the assumption that parent rank does not influence child rank once the child's education rank bin is known.

Each of these sets of CDFs defines a parent-child conditional expectation function, from which we can calculate scalar statistics of mobility. The bounds on $p_{25}$, the expected child rank at the 25th percentile of the parent distribution, are $[36.6, 39.1]$. The comparable unconstrained bounds from Table 1 are $[22.2, 54.5]$, highlighting the fact that censoring in the child distribution poses considerably less concern than censoring in the parent distribution, at least for this cohort. The bounds on the rank-rank gradient with censored child data are $[0.54, 0.64]$, compared to $[0.44, 0.66]$, again considerably more narrow. The bounds on $\mu_0^{50}$ are more similar, because as discussed above, $\mu_0^{50}$ is only minimally affected by censoring in the bottom bin of the parent distribution. We obtained $[37.0, 39.1]$ when we accounted for censoring in the parent distribution, and $[37.5, 39.0]$ when we account for censoring in the child distribution. These results show that for India, censoring in the child rank distribution is a second-order concern to censoring in the parent rank distribution, because the censoring problem is largest for the bottom bin in the parent education distribution.

Finally, because we have additional data on children, we can estimate the shape of the child CDF within parent-child education bins using rank data from other outcome variables. If children's opportunities are indeed perfectly ordered in parent ranks *within* parent-child education bins (as would be implied by the CDFs that minimize mobility in Figure C2), we would expect to see this pattern across other outcomes. Figure C3 tests this hypothesis using children's wage data. To generate this figure, we calculate children's ranks first according to education, and then according to wages within each education bin.[64] If parent education strongly predicted child wages within each education bin, we should see a graph like Panel A of Figure C2. Instead, the graph is much closer to Panel B of Figure C2; there is some additional curvature in some bins, but the distribution of child CDFs is strikingly close to that in Panel A of Figure C2. This result suggests that our assumption in the body of the paper that sons' ranks are approximated by the bin midpoint is unlikely to bias our results significantly.

Note that there is no comparable exercise that we can conduct to improve upon the situation when parent ranks are interval censored, because we have no information on parents other than their education. The closest we can come to this is by observing the parent-child rank distribution in countries with more granular parent ranks, as we did in Section IV.D. The results in that section suggest that interval censoring of parent ranks can indeed mask important features of the mobility distribution.

In this section, we have shown that (i) interval censoring in children's rank data is less consequential for mobility estimates than interval censoring in parent rank data; and (ii) our best estimates suggest that assigning children the midpoint rank in their rank bin is a good approximation of their true ranks. We expect the first of these results to hold for the parent-child education distributions in other developing countries and in any other context where the largest censoring problems are in the bottom parent bin. Because the second result can be estimated directly, it may be worth estimating in each new context. Developing a

[64]We limit the sample to the 50% of children who report wages. Results are similar if we use household income for all children. As discussed in Section II.A, both measures are imperfect.

method to solve for censored rank data in the parent and child distributions simultaneously would be valuable, but is left for future research.

**Figure C1**

Observed Son CDF by Father Education(1960-69 Birth Cohort)

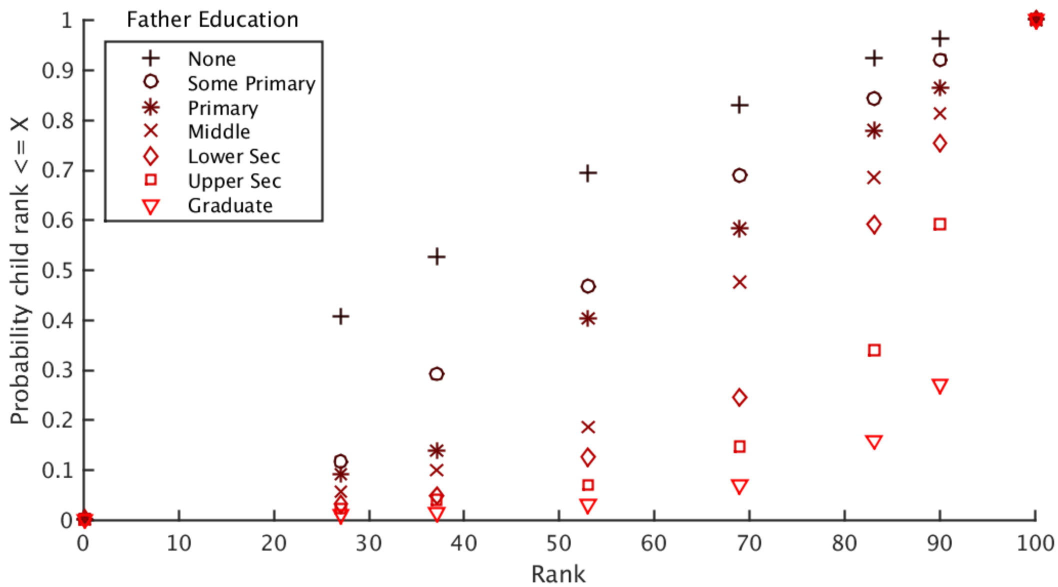
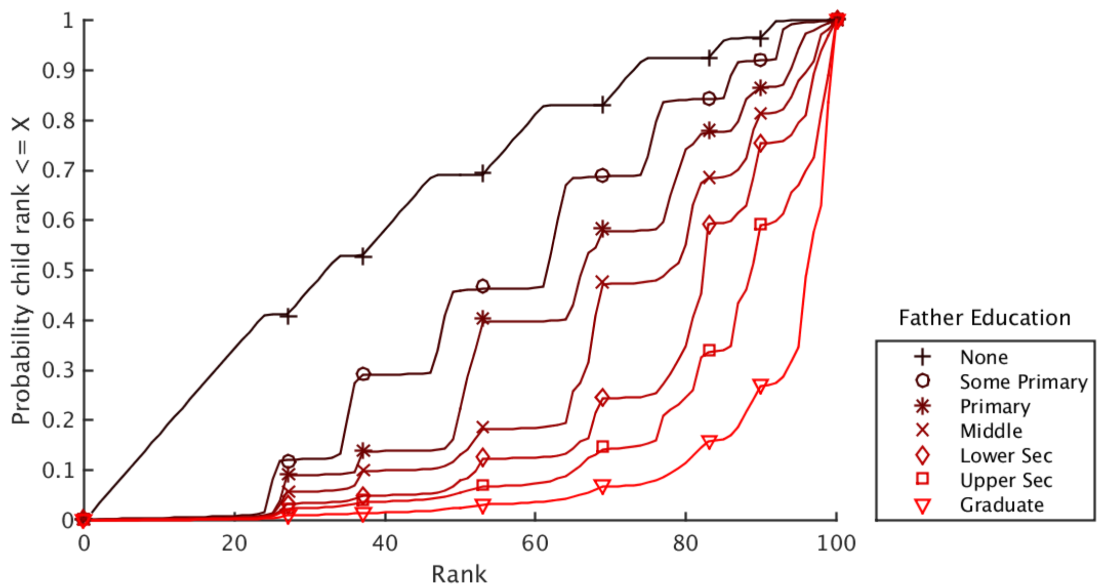



Figure C1 plots the observed points on the child CDF for each parent education group, using data from the Indian 1960-69 birth cohort and their fathers. Each marker type represents a different type of father, classified by education. For each father type, the graph shows a child's probability of attaining a given rank percentile in the child education rank distribution.

**Figure C2**

Bounds on Feasible Son CDFs by Father Education (1960-69 Birth Cohort)

Panel A: Lowest Feasible Mobility




Panel B: Highest Feasible Mobility

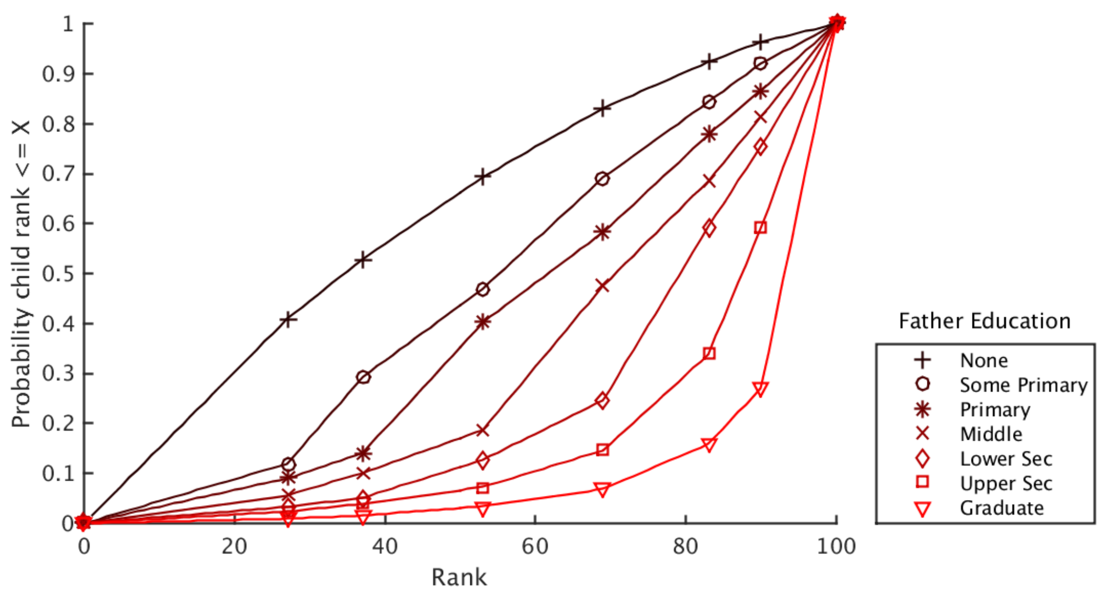



Figure C2 shows bounds on the child education rank cumulative distribution function, separately for each father education group. The lines index father types. Each point on a line shows the probability that a child of a given father type obtains an education rank less than or equal to the value on the X axis in the national education distribution. The large markers show the points observed in the data (corresponding to Figure C1).

**Figure C3**

Son Outcome Rank CDF by Father Education (1960-69 Birth Cohort)

Joint Education/Wage Estimates

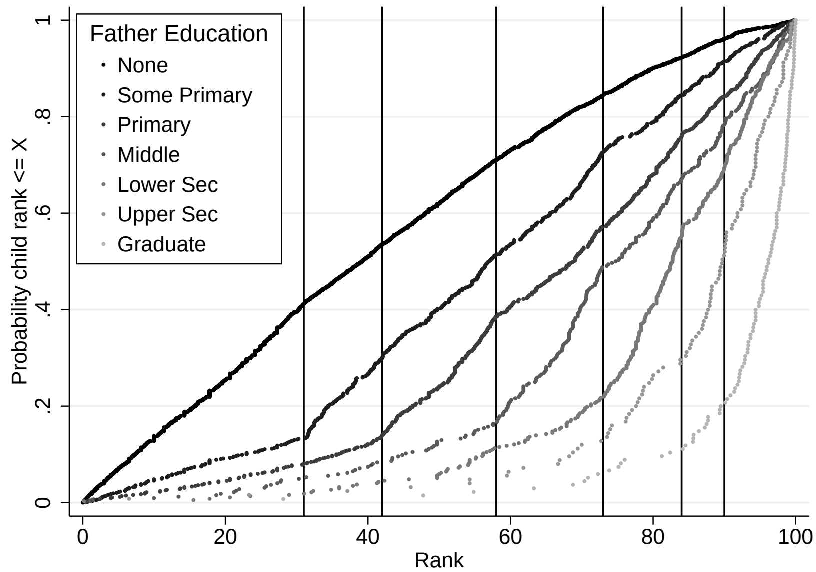



Figure C3 plots separate son rank cumulative distribution functions separately for each father education group, for sons born in the 1960s in India. Sons are ranked first in terms of education, and then in terms of wages. Data are from the IHDS 2011-12. Sons not reporting wages are dropped. For each father type, the graph shows a child's probability of attaining a given rank percentile in the child education rank distribution.