Estimating Intergenerational Mobility in Developing Countries: New Methods and Evidence from India

Sam Asher, World Bank
Paul Novosad, Dartmouth College
Charlie Rafkin, MIT

DEC Policy Research Talk

June 13, 2019

Outline

Preview

Motivation and Context

Estimating IGM in Developing Countries

Data

Results: Intergenerational Mobility in India

Understanding local growth and development

- Broader agenda examines determinants of growth and structural transformation in developing countries
- Explore within-country distribution of economic activity
- Assemble high spatial resolution administrative datasets covering every business and individual in India

Understanding local growth and development

- Broader agenda examines determinants of growth and structural transformation in developing countries
- Explore within-country distribution of economic activity
- Assemble high spatial resolution administrative datasets covering every business and individual in India
- Different papers testing the role of:
  - Natural resources
  - Local politics
  - Infrastructure (roads, canals)
  - Urban-rural spillovers

Motivation

- WDR 2006: Equity and Development starts with a comparison of expected life outcomes for two children in South Africa
  - Nthabiseng is black, poor parents, rural
  - Pieter is white, rich parents, urban
- Then only three pages dedicated to intergenerational mobility, due to constraints on data and methods
- Since then: much progress in rich countries, much less in developing countries
- Today's talk:
  - Methods to generate estimates of mobility that are valid for comparisons across groups, time, geography
  - Application to India: Ram vs Rahim

Preview of Presentation

Original Goal: Estimate intergenerational mobility in India, across subgroups and across time (fathers and sons mostly, some new results from women's data)

Challenge: Changing education distribution, coarse measurement of education

Solution: Partial Identification of expected child rank given parent rank

Some Results:

- No change in mobility from pre-liberalization to present
- Scheduled Castes and Tribes doing better, Muslims doing worse
- Substantial geographic variation

Other Applications of These Methods:

- Rising Mortality in the U.S.
- Fertility, Marriage, Bond Ratings, etc..

Outline

Preview

Motivation and Context

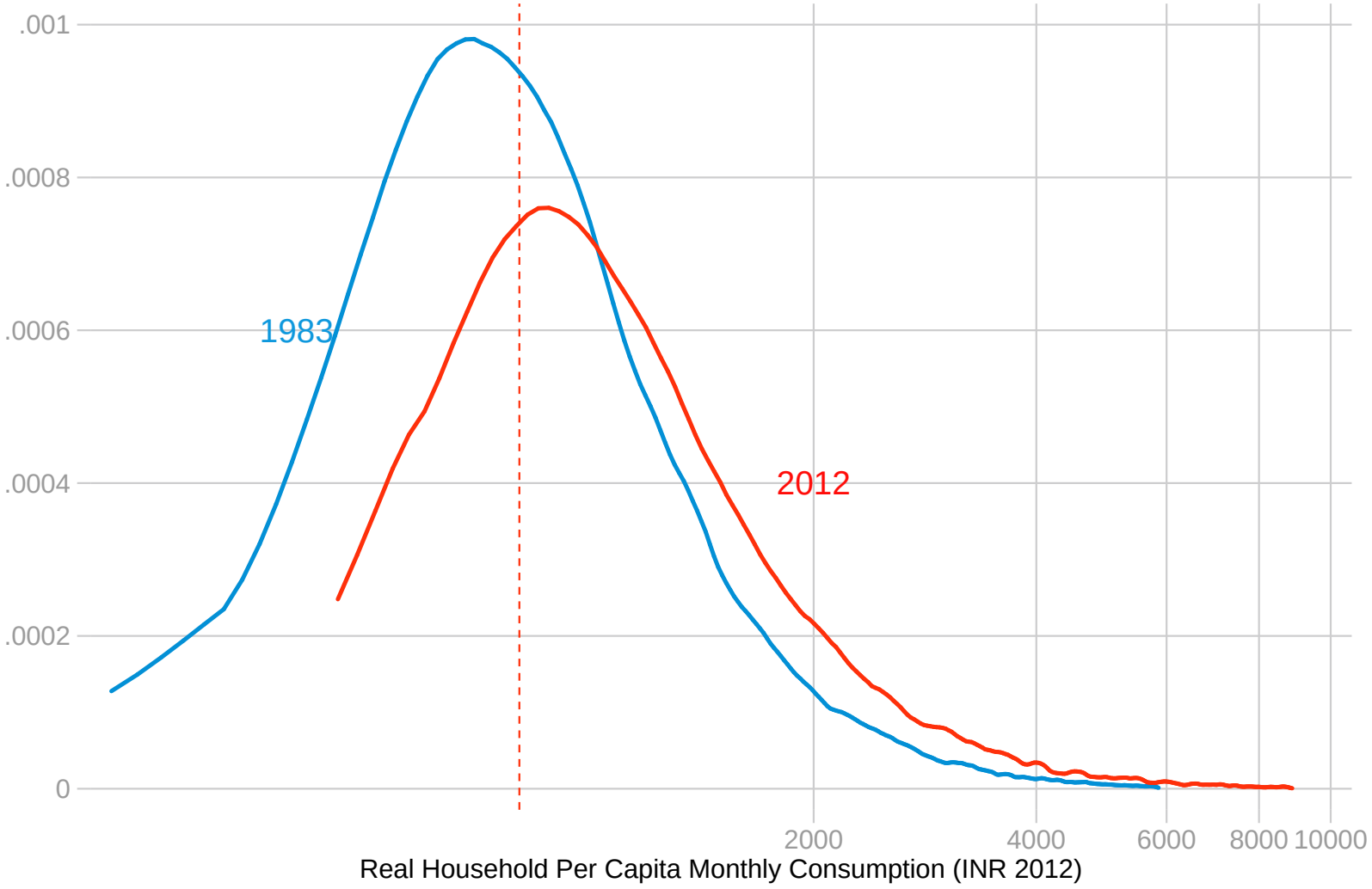
Estimating IGM in Developing Countries

Data

Results: Intergenerational Mobility in India

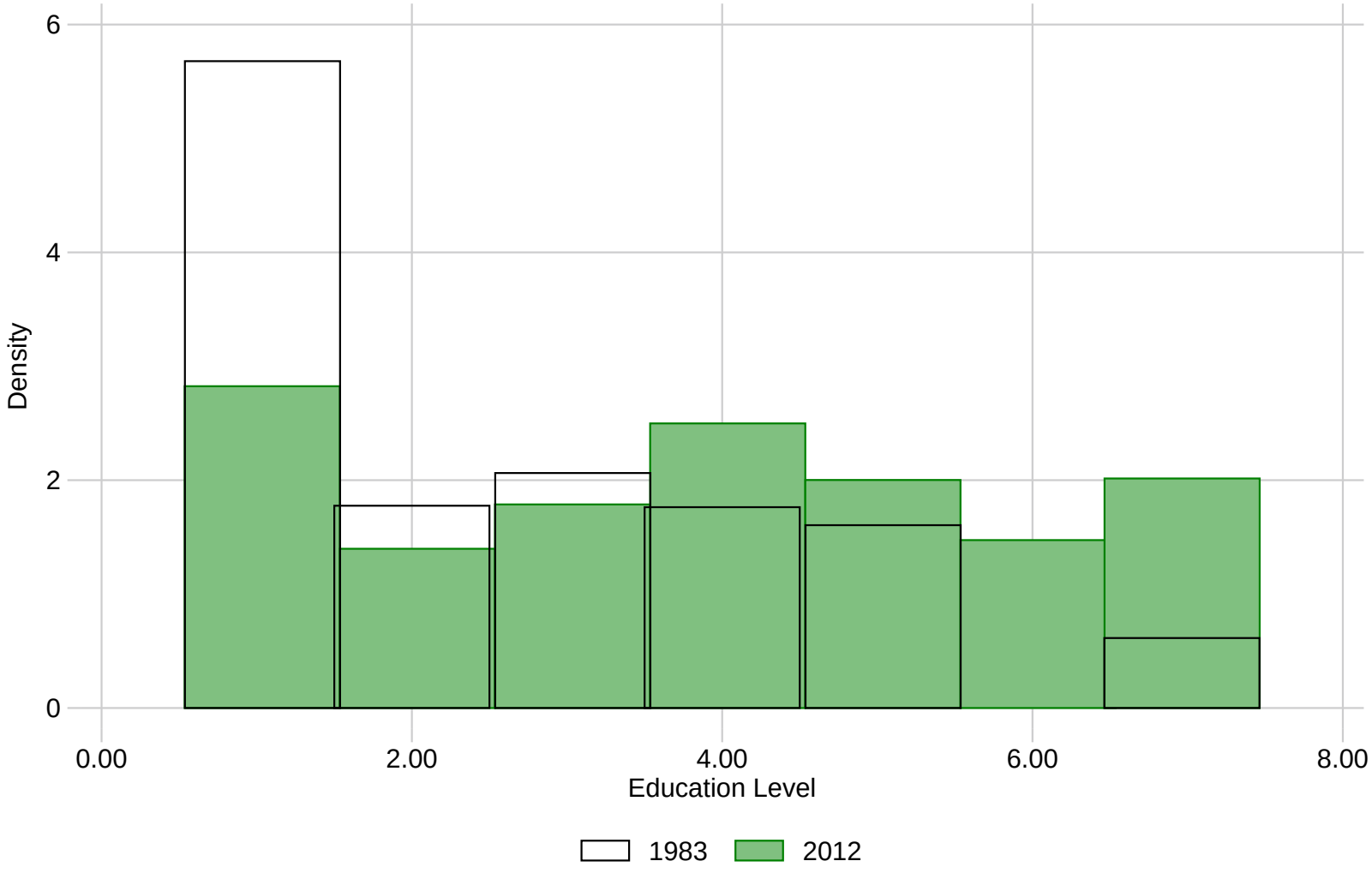
Motivation

The last three decades have been an economic miracle in India



Motivation

The last three decades have been an economic miracle in India



Two stories about modern India:

1. Markets overturn old hierarchies
   - Caste identities less important in cities
   - Political mobilization of marginalized groups
   - Growing equality of opportunity?

The New York Times

Scaling Caste Walls With Capitalism's Ladders in India

By LYDIA POLGREEN DEC. 21, 2011

THE TIMES OF INDIA

The unexpected rise of dalit millionaires

July 31, 2011, 2:57 AM IST SA Aiyar in Swaminomics | India | TOI

Two stories about modern India:

2. Persistence of Traditional Social Structure
   - Punishment of norm violators, even in cities
   - Identity-based politics
   - Persistent inequality

The Economist explains

Why caste still matters in India

Its importance has diminished in the world of work, but persists in society, particularly where marriage is concerned

CNN World +

India's caste system: Outlawed, but still omnipresent

By Ravi Agrawal

Updated 9:27 PM ET, Tue February 23, 2016

Research Agenda

Intergenerational Mobility: the dependence of children's social status on their parents' social status

- A measure of equality of opportunity
- Implications for income inequality, but can move independently
- A central political issue in many countries, including U.S., and esp. subgroup mobility
  - Affirmative action, tolerance of inequality and POUM (Benabou and Ok 2001), criminal politics in India (Vaishnav 2017)

Outline

Preview

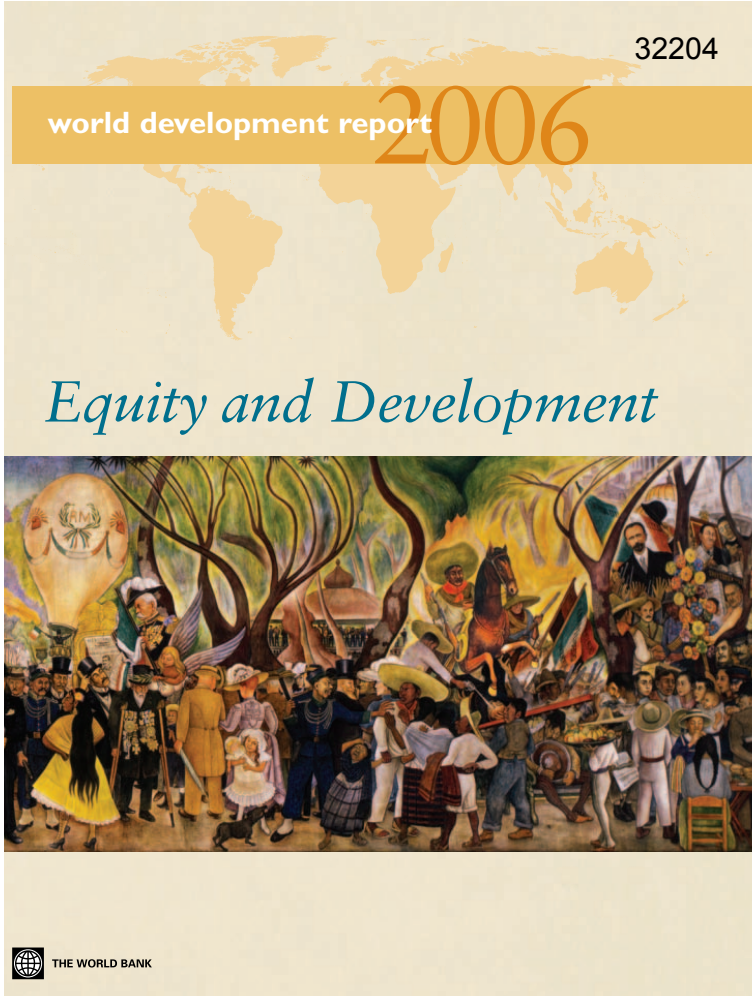
Motivation and Context

Estimating IGM in Developing Countries

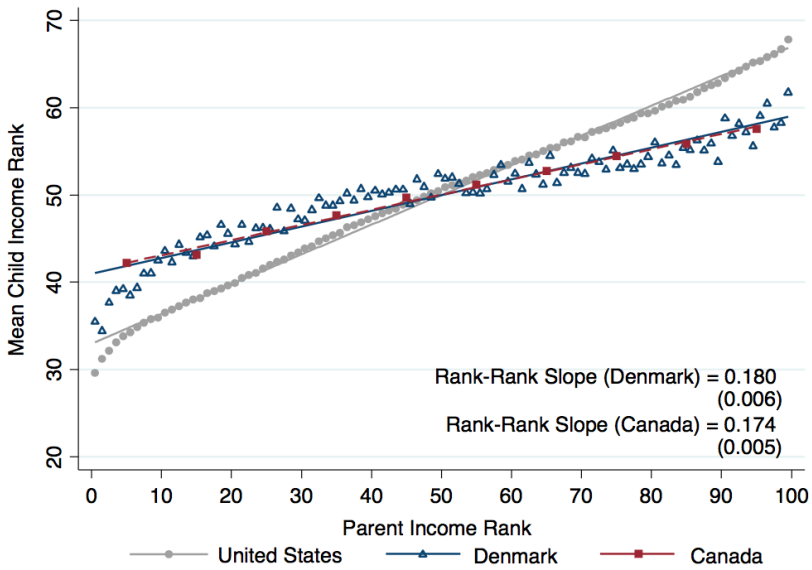
Data

Results: Intergenerational Mobility in India

Standing on the Shoulders of Giants



Measurement of Mobility (Chetty et al. 2014)



- Common measures:
  - Absolute Upward Mobility (p_{25})
  - Absolute Downward Mobility (p_{75})
  - Rank-Rank Gradient (β)

Estimating IM in Developing Countries: Challenges

- Matched parent-child income data is rarely available
  - And almost never at same age $\rightarrow$ life cycle bias
- Matched parent-child income data is unreliable
  - Key issue: measurement error $\rightarrow$ higher measured mobility
  - Measuring rural income is hard
  - How to attribute household income to coresident parents/children?
- Most studies in developing countries use education as Y
  - Arguably the best available proxy for lifetime income

Limitations of Conventional Method for IM Estimation

Standard approach:

- Linear estimation of child education (rank) on father education (rank)
- High coefficient $\rightarrow$ Low mobility

Some weaknesses of this measure:

1. Pools information from top and bottom of rank distribution

Limitations of Conventional Method for IM Estimation

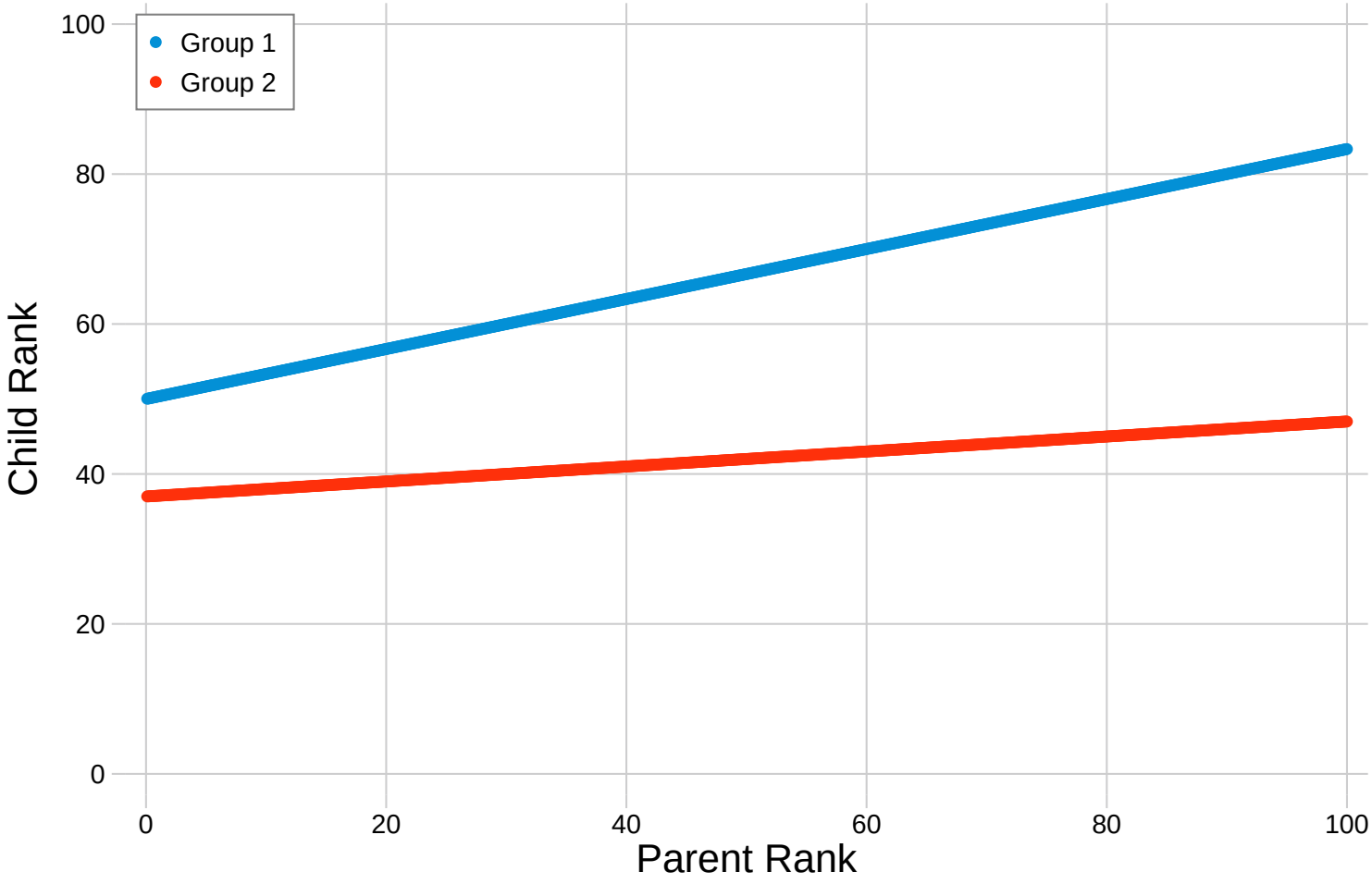
Standard approach:

- Linear estimation of child education (rank) on father education (rank)
- High coefficient $\rightarrow$ Low mobility

Some weaknesses of this measure:

1. Pools information from top and bottom of rank distribution
2. Not useful for subgroup analysis

Gradient Not Useful for Subgroup Analysis



Group 2 has a lower rank-rank gradient $\rightarrow$ more mobile?

Limitations of Conventional Method for IM Estimation

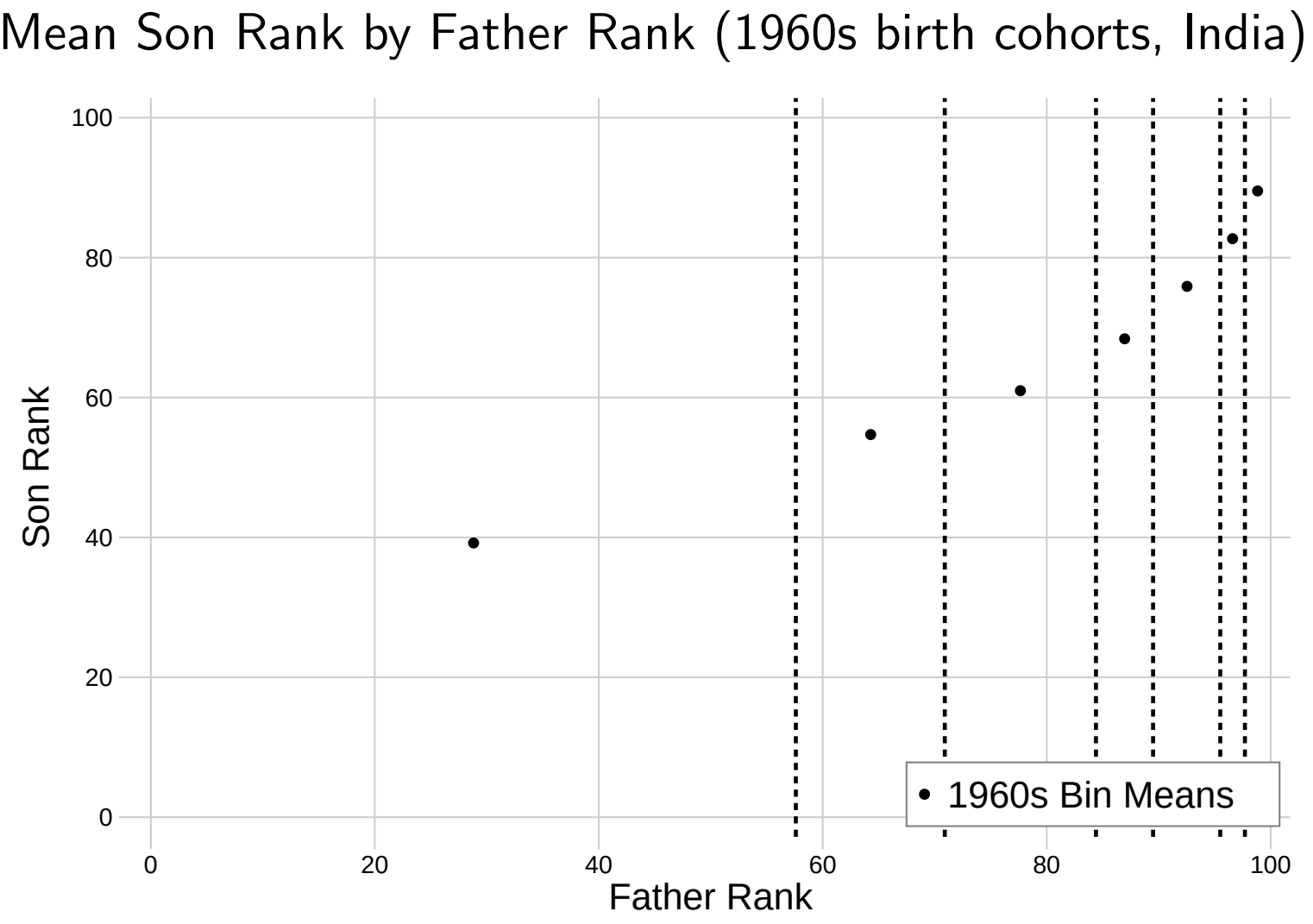
Standard approach:

- Linear estimation of child education (rank) on father education (rank)
- High coefficient $\rightarrow$ Low mobility

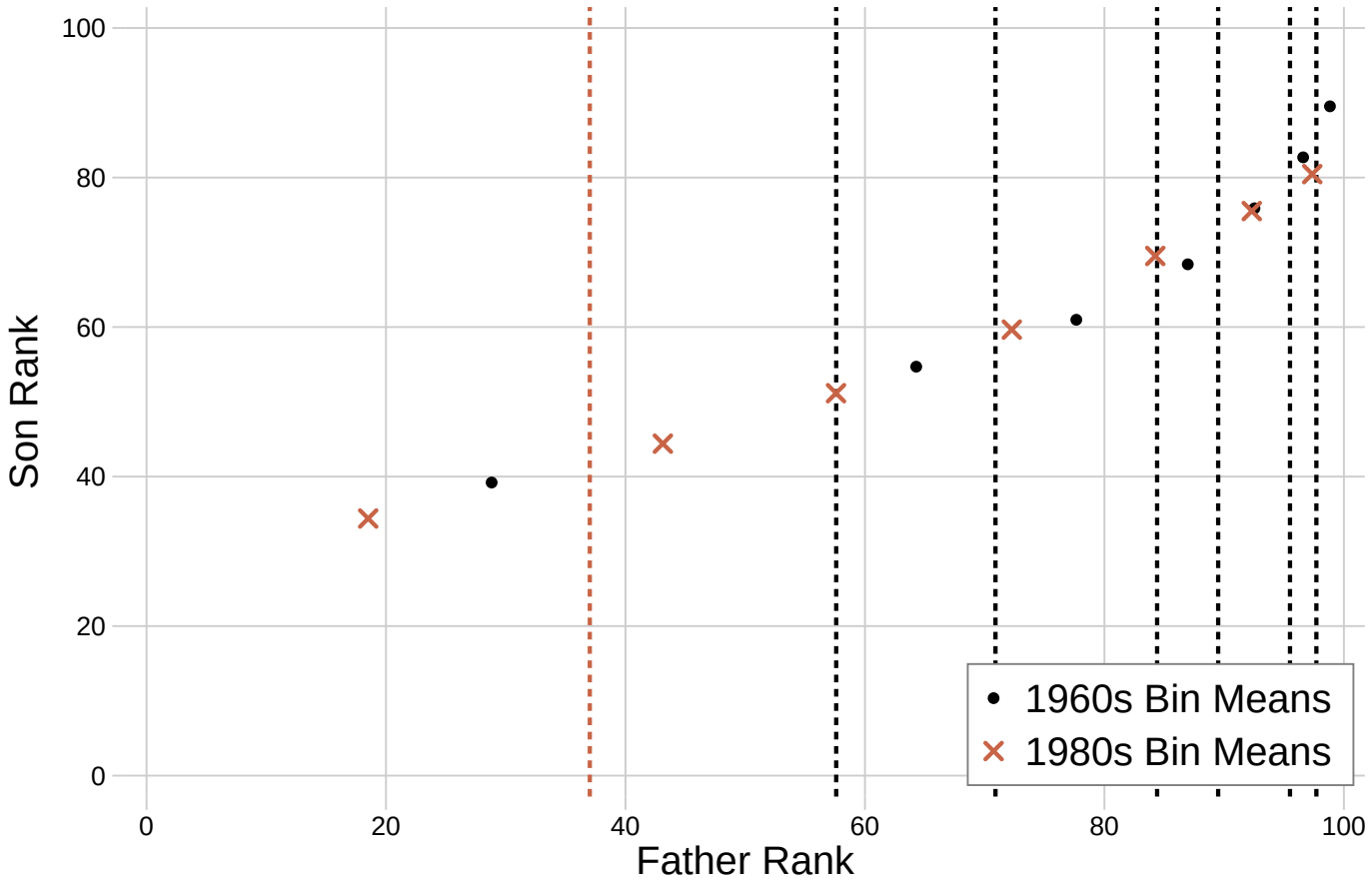
Some weaknesses of this measure:

1. Pools information from top and bottom of rank distribution
2. Not useful for subgroup analysis
3. Education is observed coarsely
   - 57% of fathers of 1960s birth cohort in India have < 2 years education
   - Internationally comparable datasets (e.g. IPUMS) use $\leq$5 ed bins

Parent-Child Education Rank Distribution, India



Comparing CEFs across time: India 1960s vs. 1980s



How should we compare the 1980 and 1960 birth cohorts?

Other Approaches in the Recent Literature

- Card et al. (2018) on educational mobility (IEM) in the 1920s
  - Definition: the 9th grade completion rate of children whose parents have 5–8 years of school
    - This is approximately $E(y > 50|x = 50)$
  - Compare this in 1980s with Chetty measure $E(y|x = 25)$

Other Approaches in the Recent Literature

- Card et al. (2018) on educational mobility (IEM) in the 1920s
  - Definition: the 9th grade completion rate of children whose parents have 5–8 years of school
    - This is approximately $E(y > 50|x = 50)$
  - Compare this in 1980s with Chetty measure $E(y|x = 25)$
- Alesina et al. (2019) on IEM in Sub-Saharan Africa
  - Definition: Probability that a child completes primary school conditional on a parent who didn't
    - This is $E(y > 52|x \in [0, 76])$ in Mozambique...

Other Approaches in the Recent Literature

- Card et al. (2018) on educational mobility (IEM) in the 1920s
  - Definition: the 9th grade completion rate of children whose parents have 5–8 years of school
    - This is approximately $E(y > 50|x = 50)$
  - Compare this in 1980s with Chetty measure $E(y|x = 25)$
- Alesina et al. (2019) on IEM in Sub-Saharan Africa
  - Definition: Probability that a child completes primary school conditional on a parent who didn't
    - This is $E(y > 52|x \in [0, 76])$ in Mozambique...
    - ... and $E(y > 18|x \in [0, 42])$ in South Africa

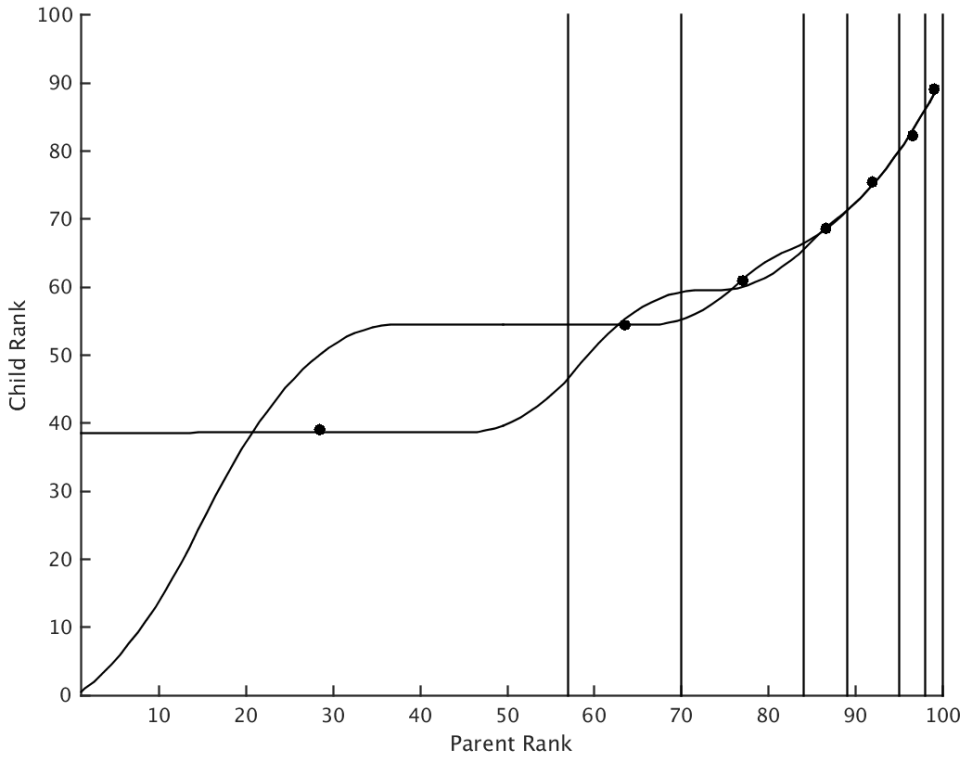
Other Approaches in the Recent Literature

- Card et al. (2018) on educational mobility (IEM) in the 1920s
  - Definition: the 9th grade completion rate of children whose parents have 5–8 years of school
    - This is approximately $E(y > 50|x = 50)$
  - Compare this in 1980s with Chetty measure $E(y|x = 25)$
- Alesina et al. (2019) on IEM in Sub-Saharan Africa
  - Definition: Probability that a child completes primary school conditional on a parent who didn't
    - This is $E(y > 52|x \in [0, 76])$ in Mozambique...
    - ... and $E(y > 18|x \in [0, 42])$ in South Africa
- Our goal: calculate $E(y|x \in [a, b])$ for any a and b

Our Strategy: A Partial Identification Approach

- **Key Idea**: Under minimal assumptions, we can *bound* the set of feasible mobility functions
- **Goal**: Conditional expectation function of child rank given parent *education percentile rank*
  - Call this $E(y|x = i)$
  - From this function, we can calculate p_{25}, p_{75}, β, and other measures of mobility
- **Problem**: Education rank X is interval censored — only observed in coarse bins
- **Solution**: Use techniques from partial identification to bound child rank (Manski and Tamer 2002; Asher, Novosad, and Rafkin 2018)

Two Candidate Father-Son CEFs: India (1960s birth cohort)



Key question: What can we say about the latent conditional expectation function?

- Both of these CEFs $Y(i)$ fit the data with zero MSE

Overview of Methods

- Assumption 1: There exists a latent education rank, observed in coarse intervals

Overview of Methods

- Assumption 1: There exists a latent education rank, observed in coarse intervals
- Arises out of the most standard human capital investment model (e.g. Card 1999, Card et al. 2018)
  - Schooling choice determined by heterogeneous cost and benefit shifters
  - Model suggests a continuous optimal level of schooling E for each individual
  - Individuals complete the last year with positive expected value
- High ranked individuals within bin would advance to next level if marginal cost/benefit shifted only a little
- Note: this is a descriptive exercise
  - We are not trying to estimate causal effects of parent education

Overview of Methods

- Assume:
  1. There exists a latent education rank, observed in coarse intervals
  2. Monotonicity: Expected child rank is weakly increasing in parent rank (Dardanoni 2012)
  3. Child CEF has discrete jumps or kinks at major education boundaries only (if at all)
  4. Child rank directly observed (loosened in paper)

Overview of Methods

- Assume:
  1. There exists a latent education rank, observed in coarse intervals
  2. Monotonicity: Expected child rank is weakly increasing in parent rank (Dardanoni 2012)
  3. Child CEF has discrete jumps or kinks at major education boundaries only (if at all)
  4. Child rank directly observed (loosened in paper)

- Manski and Tamer (MT 2002)
  - Bound a CEF $E(y|x)$ with interval-censored x
  - But bounds are too wide to be meaningful in this case

Overview of Methods

- Assume:
  1. There exists a latent education rank, observed in coarse intervals
  2. Monotonicity: Expected child rank is weakly increasing in parent rank (Dardanoni 2012)
  3. Child CEF has discrete jumps or kinks at major education boundaries only (if at all)
  4. Child rank directly observed (loosened in paper)

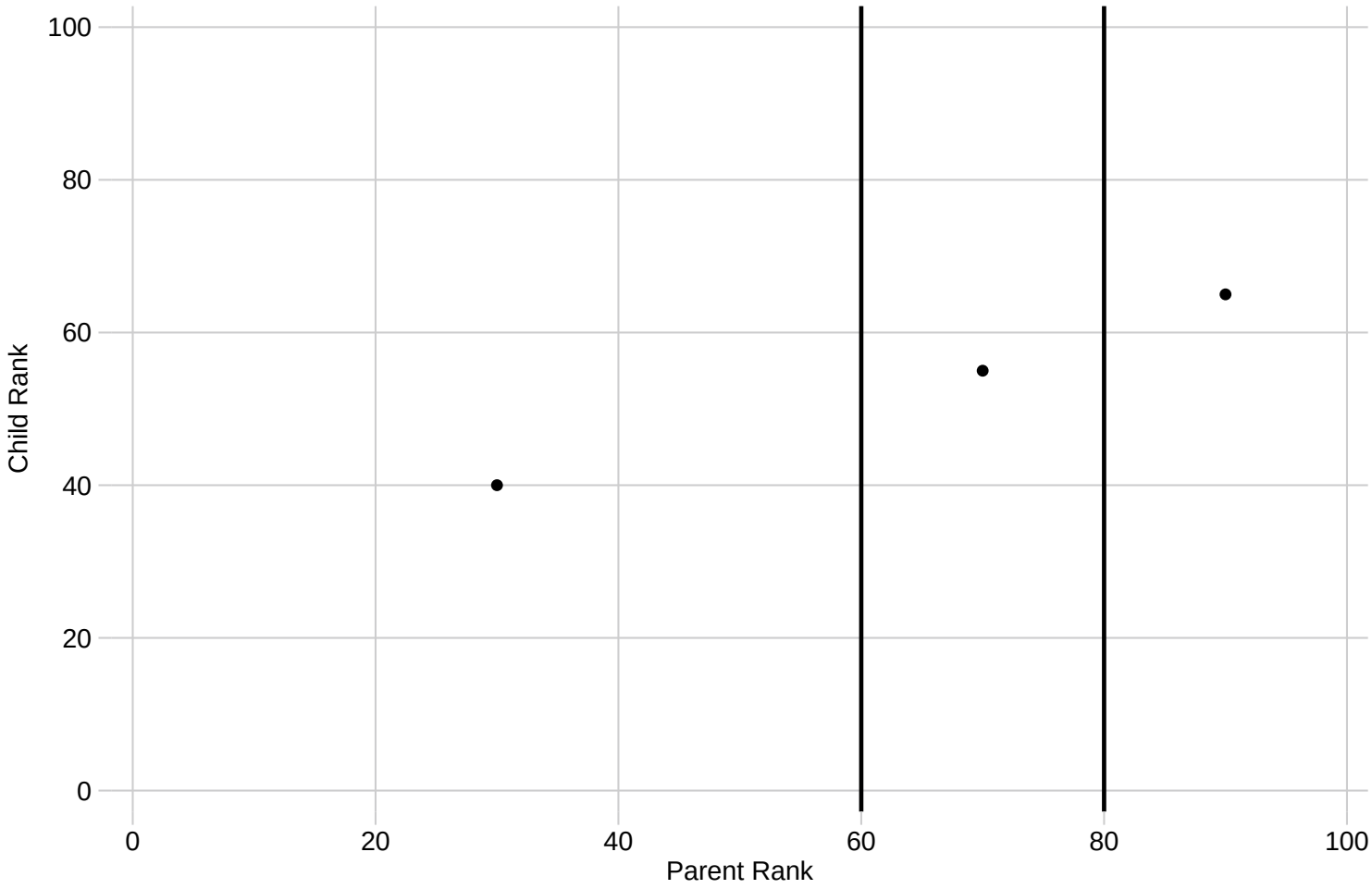
- Manski and Tamer (MT 2002)
  - Bound a CEF $E(y|x)$ with interval-censored x
  - But bounds are too wide to be meaningful in this case

- Three extensions to MT can get us useful estimates:
  1. Use knowledge of the distribution (ranks are uniform)
  2. Constrain curvature of CEF
  3. Estimate $E(y|x \in (a, b))$ instead of $E(y|x = i)$

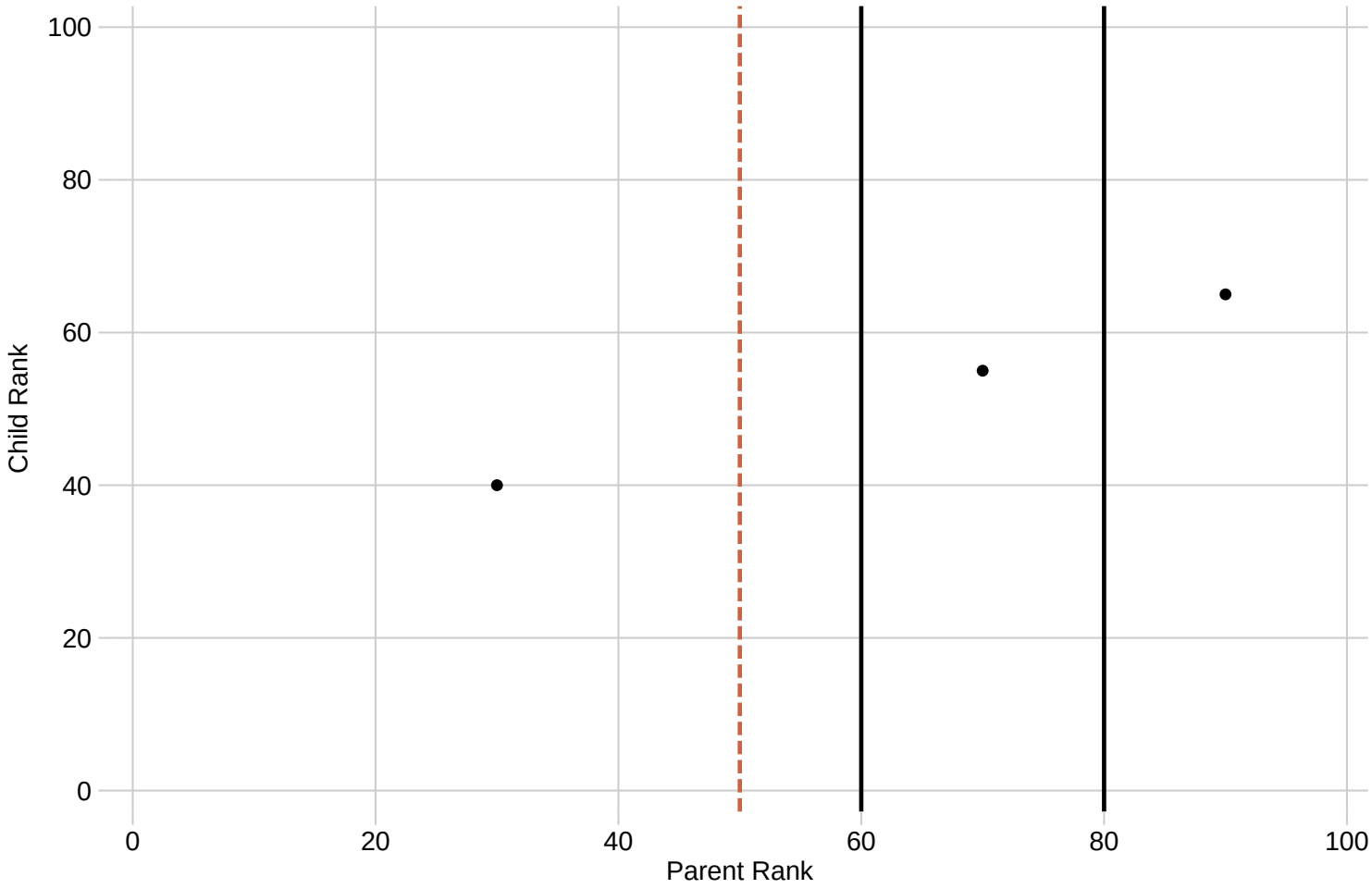
An Example for Intuition

- Assume $E(y|x \in (0, 60)) = 40$, $E(y|x \in (60, 80)) = 55$.
- We want to know $z = E(y|x \in (0, 50))$.

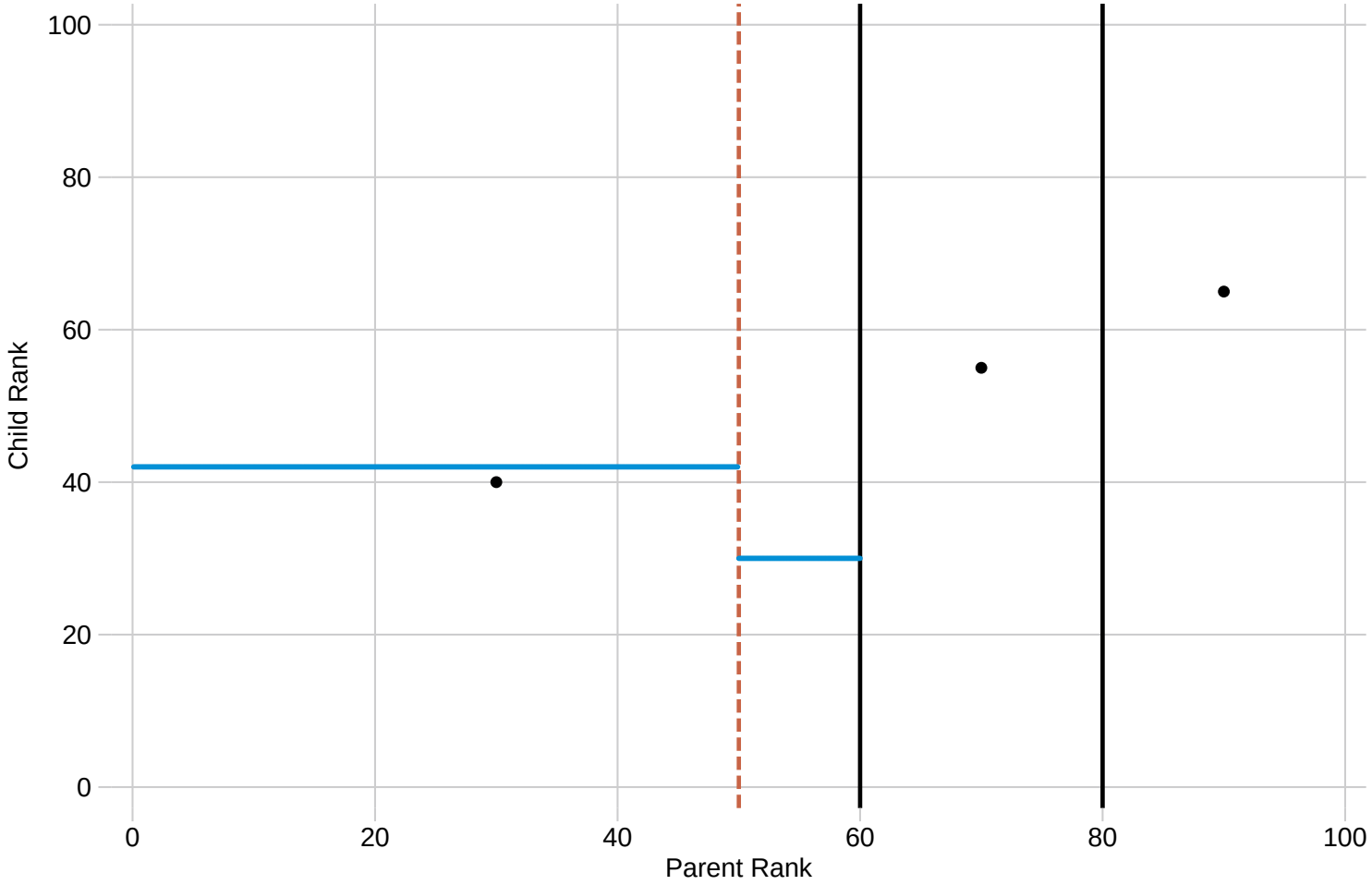
An Example for Intuition



An Example for Intuition

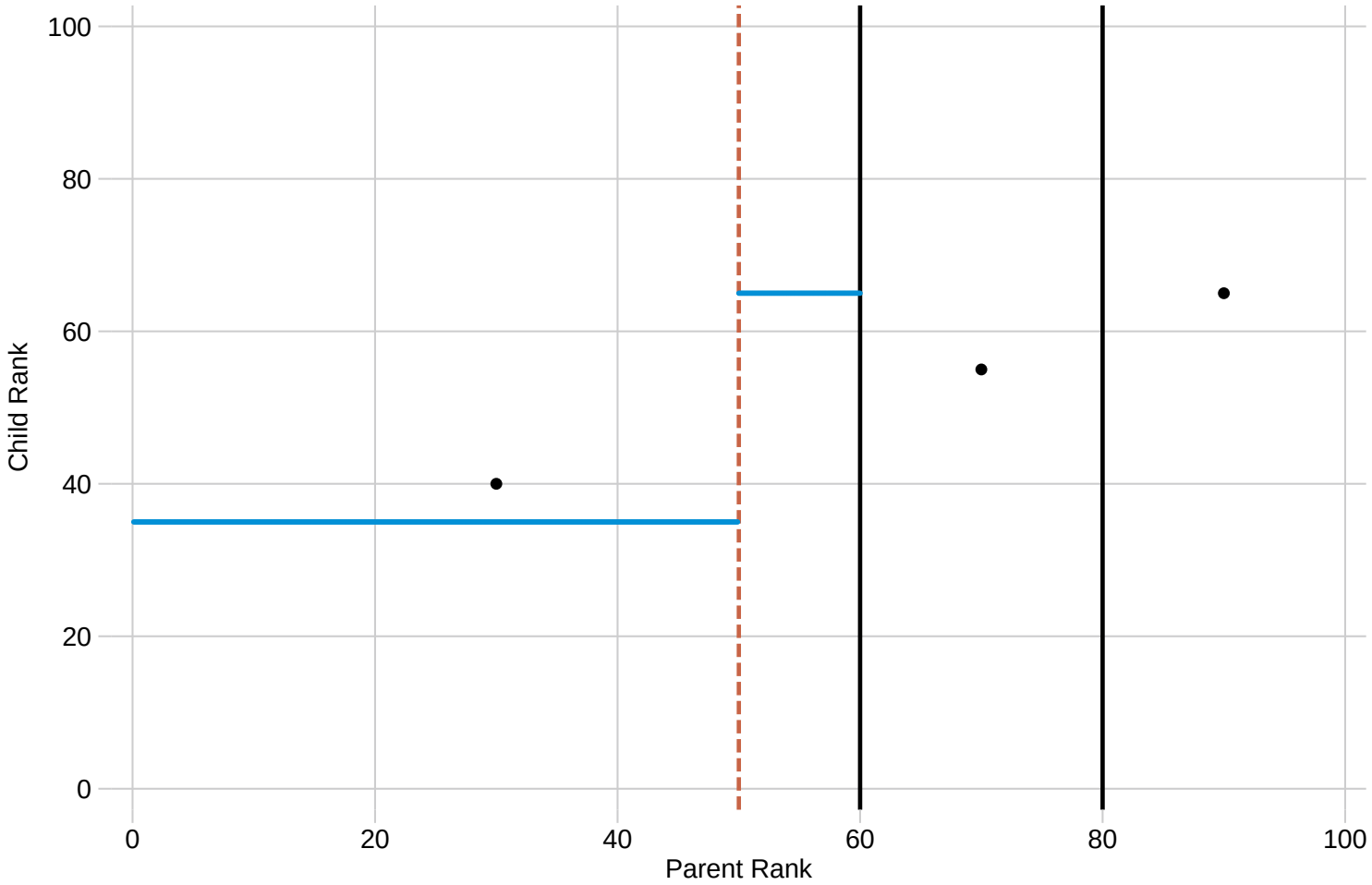


An Example for Intuition



Monotonicity within bin $\rightarrow z \leq 40$

An Example for Intuition



Monotonicity with next bin $\rightarrow z \geq 37$ — but suggests desirability of curvature constraint.

Constraining Curvature of the CEF

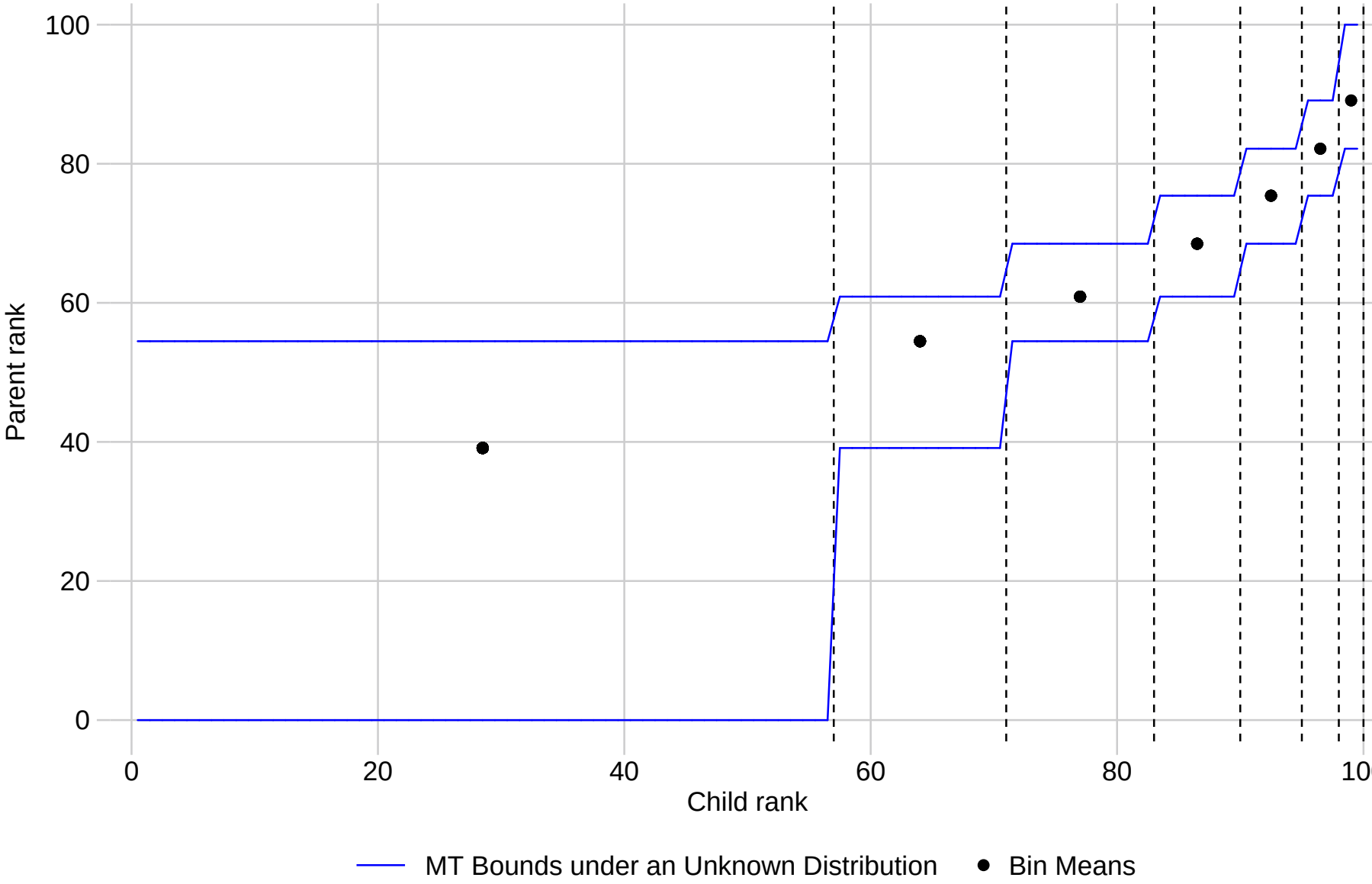
- Edges of CEFs imply large discrete changes at arbitrary rank intervals
  - These are unlikely to occur in real life
- Add assumption $Y''(x) <= \overline{C}$
- Numeric rather than analytic solution (currently) Setup
  - Matlab solver calculates bounds on discrete approximation to CEF
  - CEF approximation takes the form of 100 values for 100 integer ranks
- Permit (but do not impose) sheepskin effects by allowing discrete jumps at bin thresholds (e.g. where people attain degrees)
- Most extreme curvature constraint $\overline{C} = 0$ gives us the rank-rank gradient
  - Curvature-constrained estimation generalizes the standard approach

Choosing a Curvature Restriction

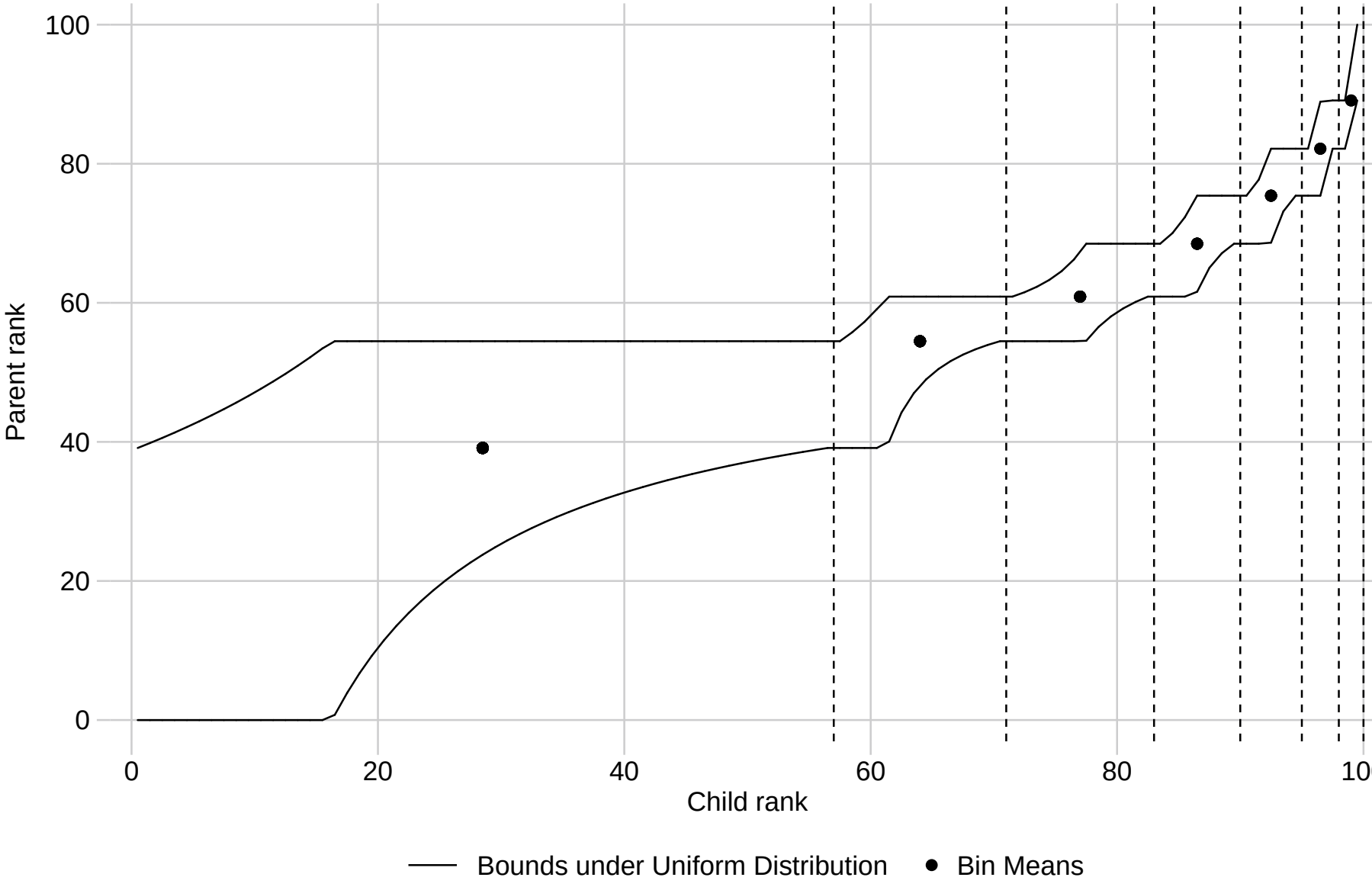
How to choose $\overline{C}$?

- We found parent-child *income* rank CEFs for USA, Denmark, Sweden, Norway Chetty et al. (2016), Bratberg et al. (2017)
- Fit splines to each of these and calculate max 2nd derivative Example Spline
- Conservative $\overline{C}$: Double the maximum value of 1.6 to get $\sim \overline{C} = 3$
- All results go through without curvature constraint

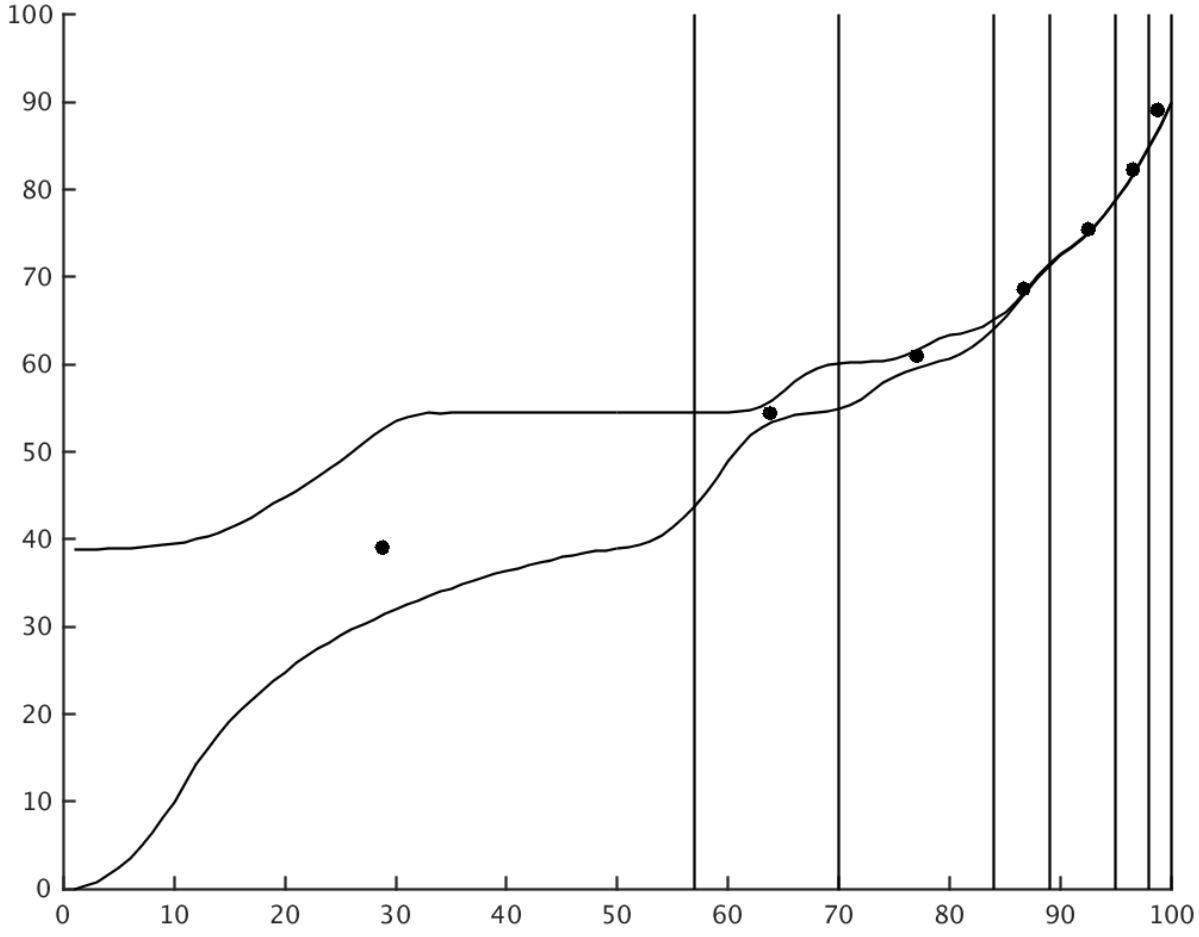
CEF Bounds under Interval Data: Manski-Tamer 2002



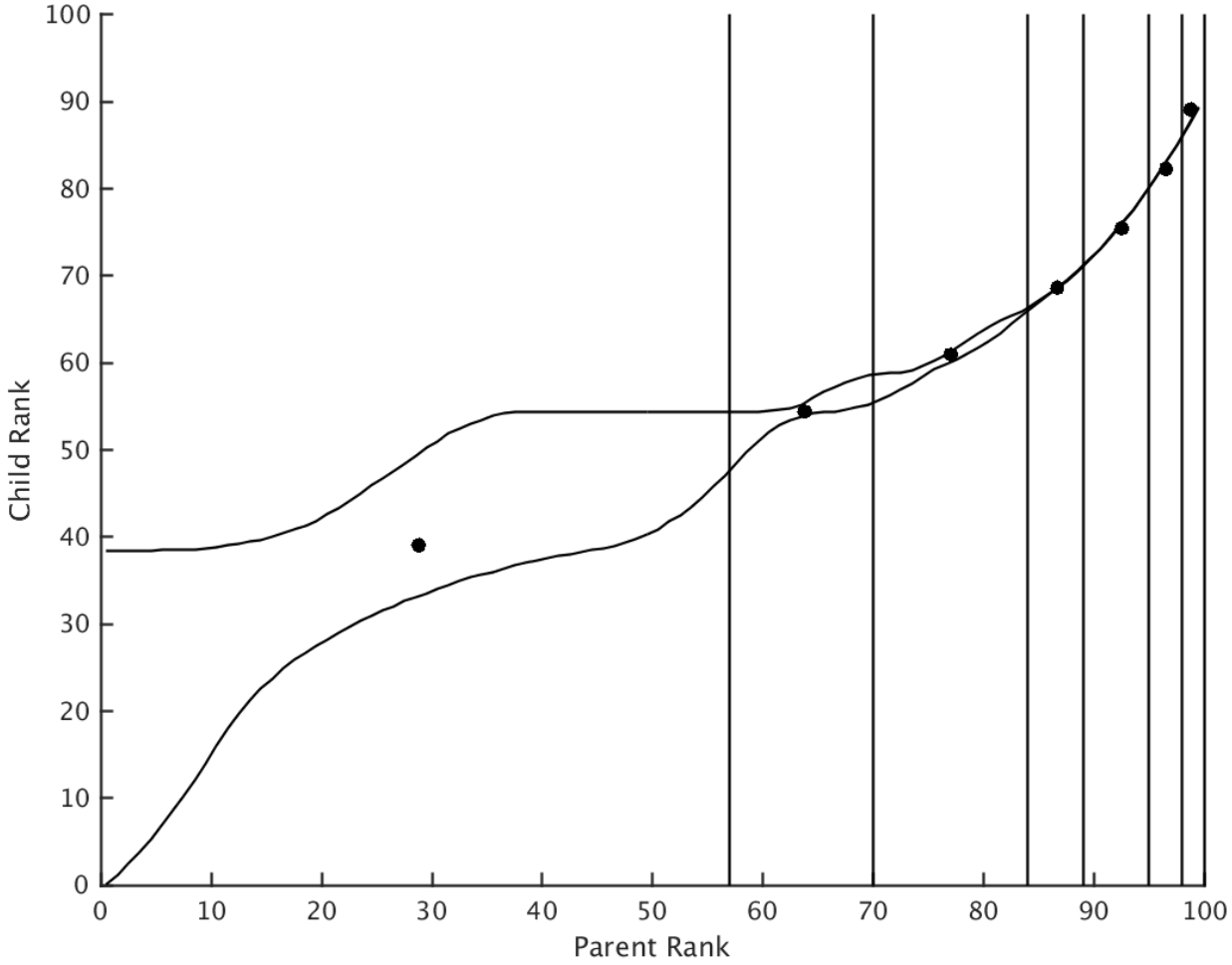
CEF Bounds under Interval Data: Uniform X



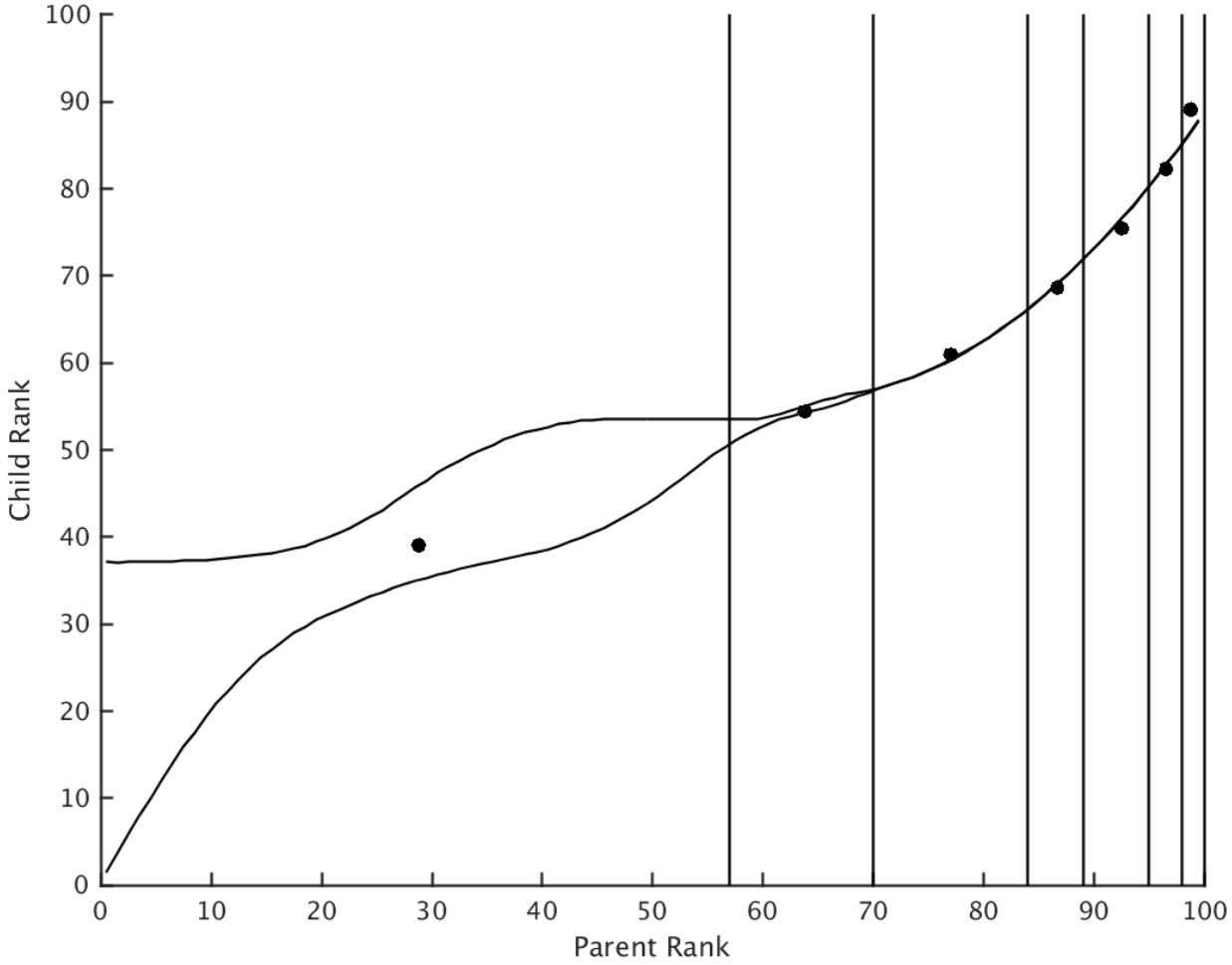
CEF Bounds under Interval Data: $\overline{C} = 20$



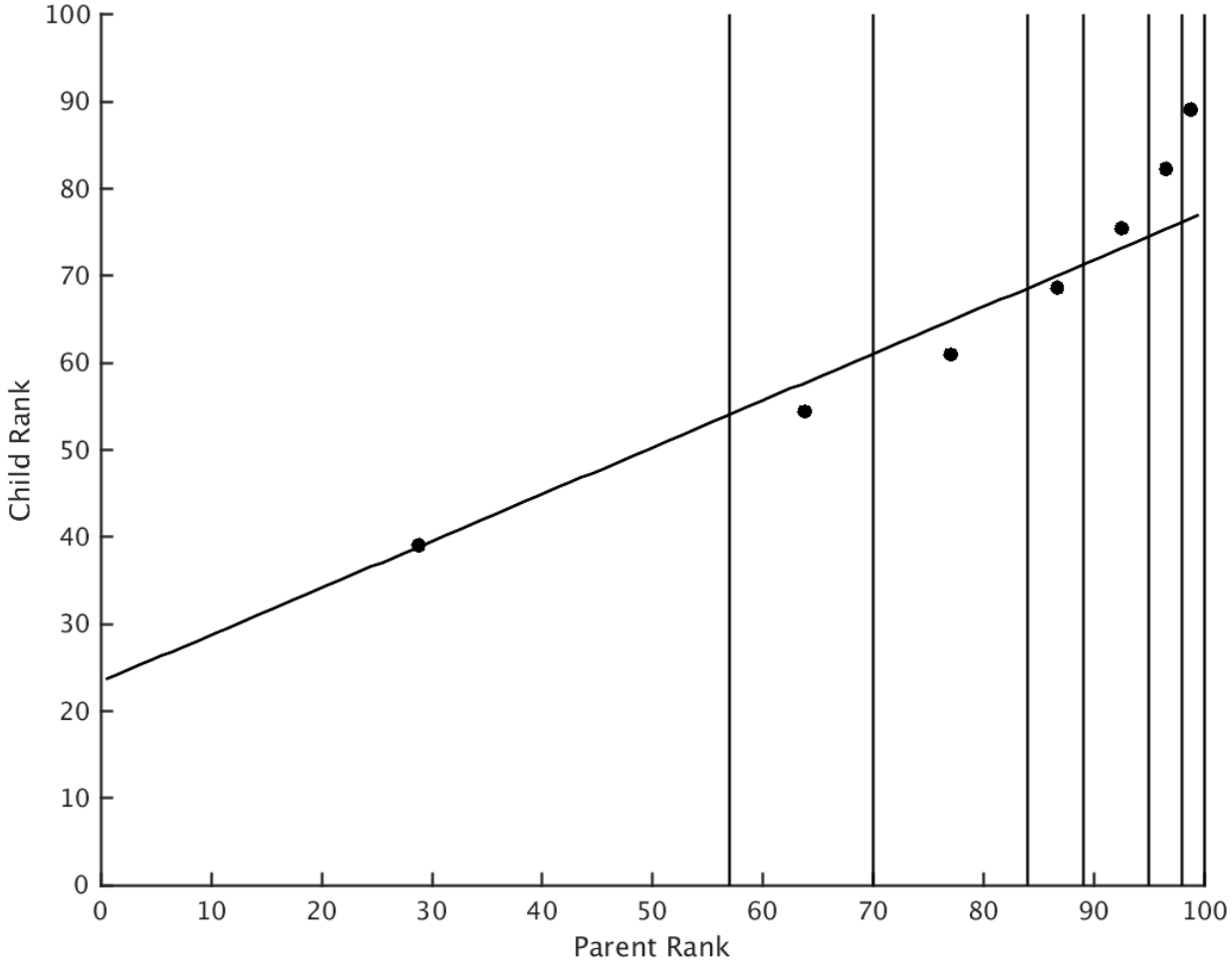
CEF Bounds under Interval Data: $\overline{C} = 10$



CEF Bounds under Interval Data: $\overline{C} = 3$



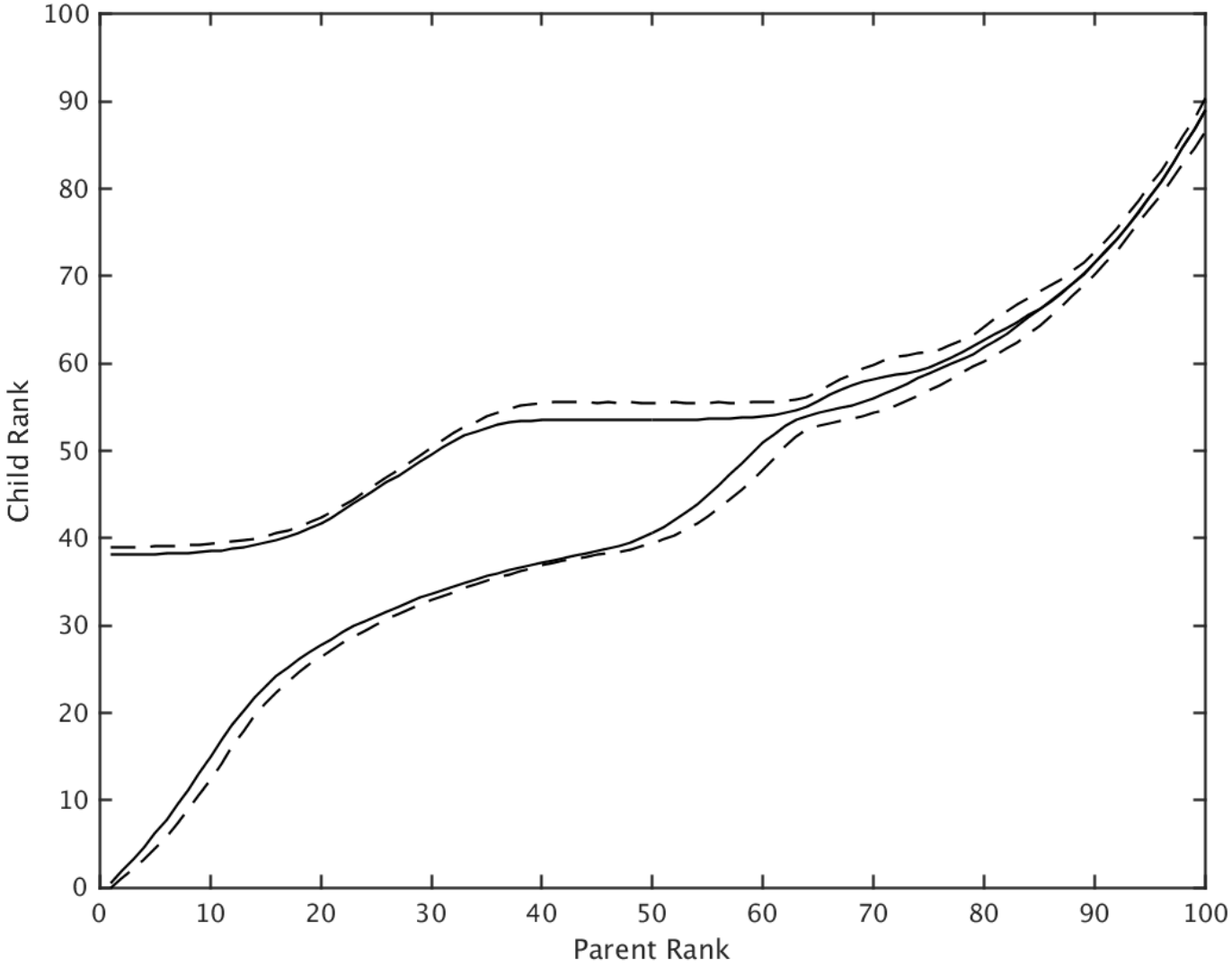
CEF Bounds under Interval Data: $\overline{C} = 0$



Some Additional Details

- Validate this approach using fully supported income rank CEF from Denmark (Bratberg et al. 2017)
  - Interval censor the data using education bin boundaries, and recover bounds on the original CEF Simulation
- Can assume curvature constraint and loosen monotonicity assumption
  - Allows estimation of interval-censored function even without monotonicity assumption
- Get standard errors through bootstrap

CEF Bounds With Bootstrap Errors: $\overline{C} = 3$



Introducing μ_a^b

- Same structure allows us to bound any function of the CEF
- We focus on one function of interest $\mu_a^b = E(Y|x \in (a, b))$
  - *e.g.*, μ_0^{25}: expected child rank, given a parent in the bottom 25%
- a and b can be arbitrary and unrelated to actual bin boundaries

- Bounds on μ_a^b:
  - Trivially point estimated if a and b are boundaries in the data
  - Tightest when bin boundaries are close to a, b

- We will use μ_0^{50} and μ_{50}^{100}
  - μ_0^{50} is a measure of upward mobility analogous to p_{25} from Chetty et al.
    - p_{25} is the expected rank of a child born to the median parent in the bottom half
    - μ_0^{50} is the expected rank of a child born to any parent in the bottom half
  - If the CEF is linear, $\mu_0^{50} = p_{25}$

Mobility Statistics: 1960s birth cohort

Bounds on key mobility statistics $\overline{C} = 3$:

Gradient β:	[0.45, 0.63]
Abs. Mobility p_{25}:	[31.0, 46.0]
Interval Mobility μ_0^{50}:	[36.5, 38.5]

Why are μ_0^{50} bounds so much tighter?

- Bin 1 has fathers in ranks 1-57:
  - μ_0^{57} is point identified – we observe it in the data
- μ_0^X is mean of $\mu_0^1, \mu_1^2, ..., \mu_{X-1}^X$
- Given μ_0^{57} and monotonicity, narrow set of possible values of μ_0^{50}.
- In paper: proof of analytical bounds for μ_a^b given interval data

Comparison with Other Approaches

Other approaches to dealing with coarse data

- Focus on groups for whom education has not changed very much
  - Useful, but limiting

Comparison with Other Approaches

Other approaches to dealing with coarse data

- Focus on groups for whom education has not changed very much
  - Useful, but limiting
- Assume linearity of CEF
  - Canonical approach in intergenerational educational mobility
  - But many fully supported CEFs are concave at bottom
  - Doesn't distinguish change at top from change at bottom
  - Identical to our approach with $\overline{C} = 0$

Comparison with Other Approaches

Other approaches to dealing with coarse data

- Focus on groups for whom education has not changed very much
  - Useful, but limiting
- Assume linearity of CEF
  - Canonical approach in intergenerational educational mobility
  - But many fully supported CEFs are concave at bottom
  - Doesn't distinguish change at top from change at bottom
  - Identical to our approach with $\overline{C} = 0$
- Randomly reassign people across bins to get same bin sizes
  - Common in the mortality / education literature
  - Used in Fair Progress (2018)
    - Concludes Ethiopia has almost perfect upward mobility

Outline

Preview

Motivation and Context

Estimating IGM in Developing Countries

Data

Results: Intergenerational Mobility in India

Data

Two Data Sources

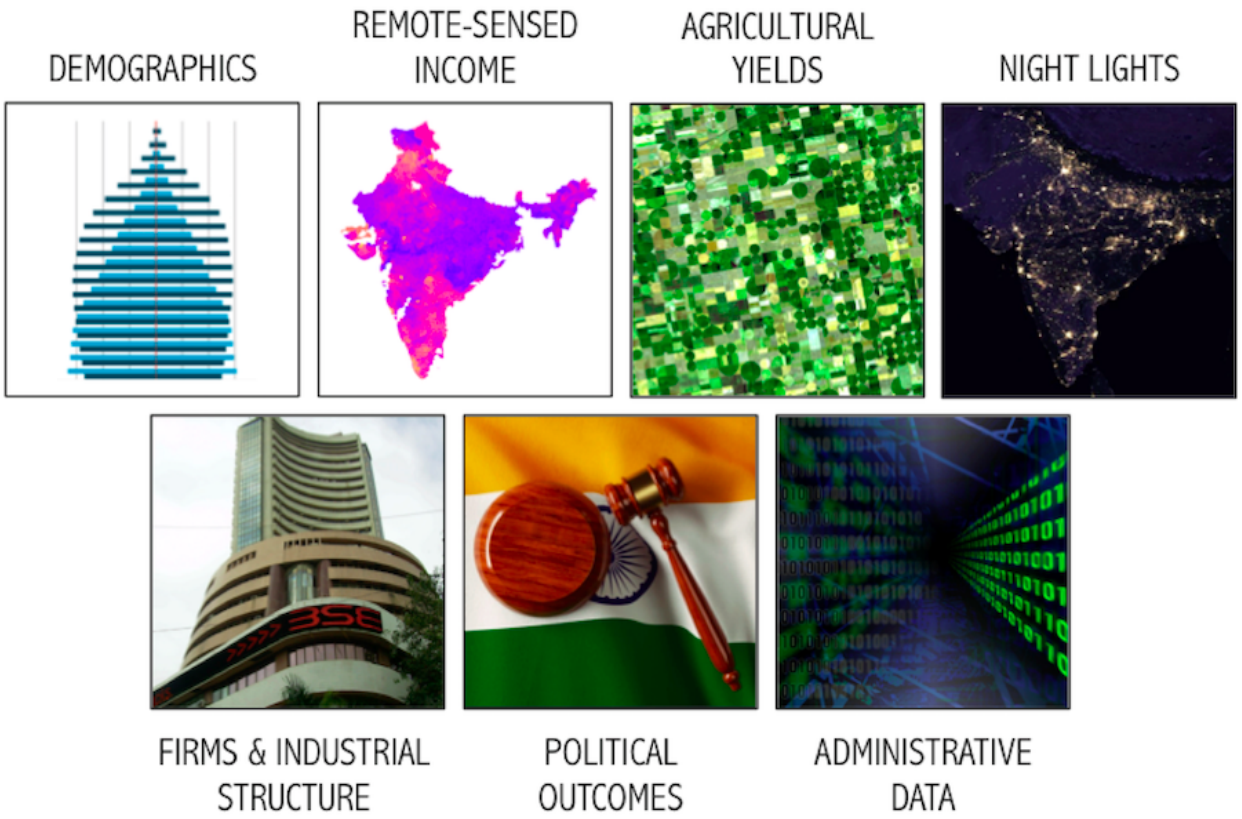
1. India Human Development Survey
   - Household survey, men are asked about fathers' education
   - Because education is static, records mobility going back 30 years
   - Subgroup characteristics: caste and religion
   - Mothers/Daughters — data limitations, some preliminary results

2. Socioeconomic and Caste Census (2012)
   - Complete enumeration of household demographics and assets
   - One of many administrative data sources used internally by government
   - 31 million sons aged 20-23 who coreside with fathers
   - Limited subgroup characteristics—but has names

The SHRUG

The Socioeconomic High-res Rural-Urban Geographic panel for India (SHRUG)

- High resolution socioeconomic aggregation of remote sensing and administrative data
- 600,000 villages and 5,000 towns from 1990 to the present
- Version 1.0 release is imminent



Outline

Preview

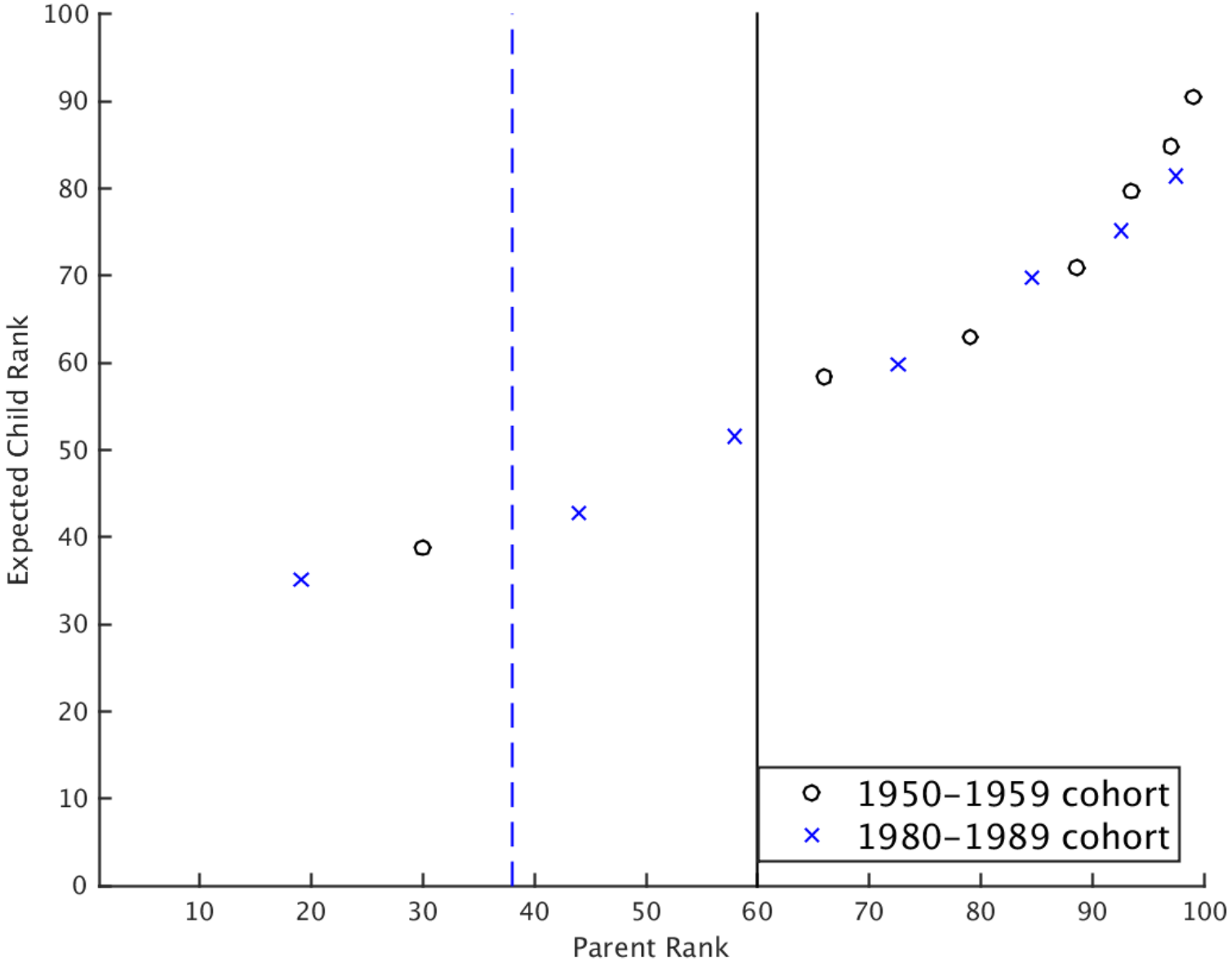
Motivation and Context

Estimating IGM in Developing Countries

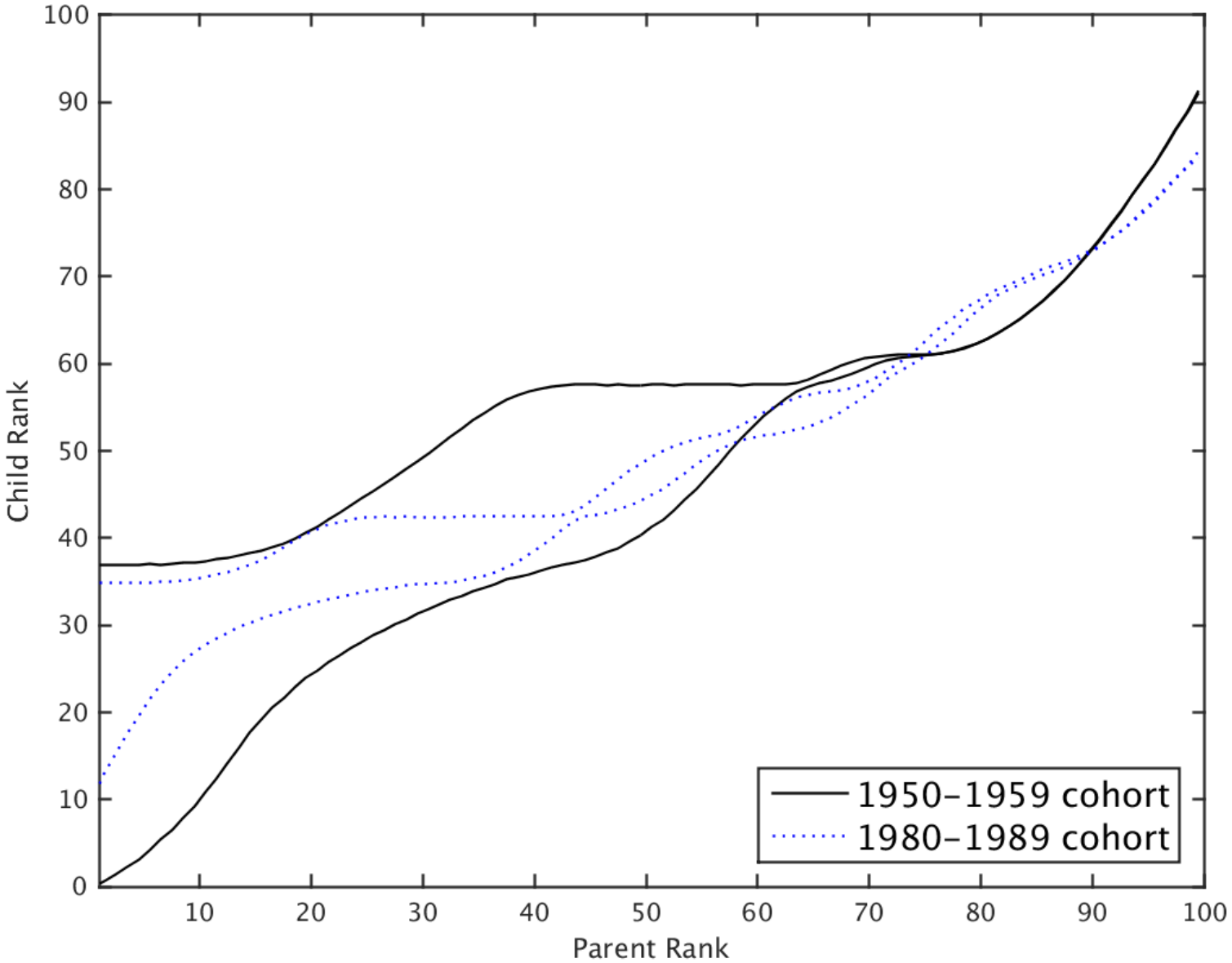
Data

Results: Intergenerational Mobility in India

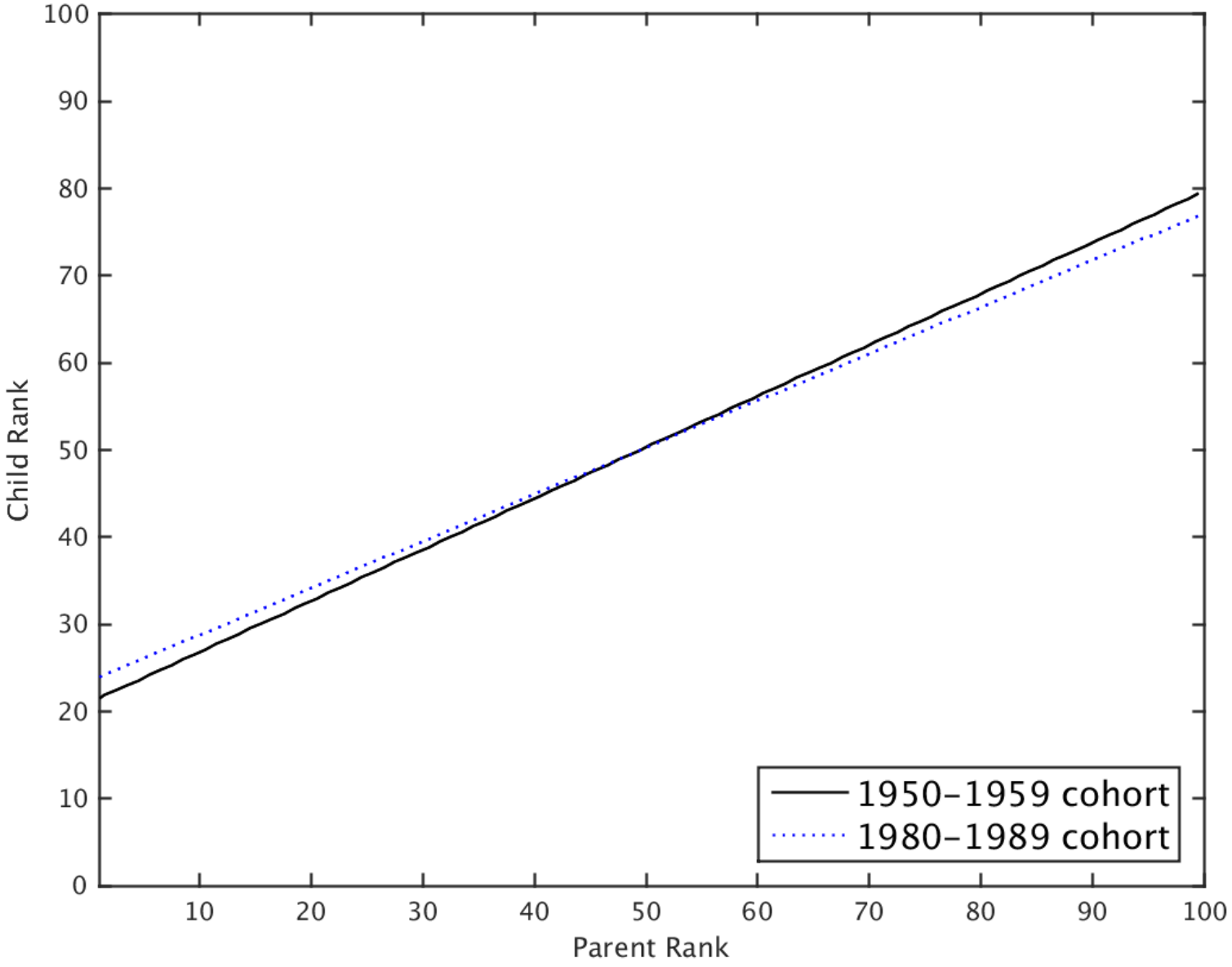
Comparing 1950s to 1980s birth cohort (raw data)



Comparing 1950s to 1980s birth cohort (preferred estimate)

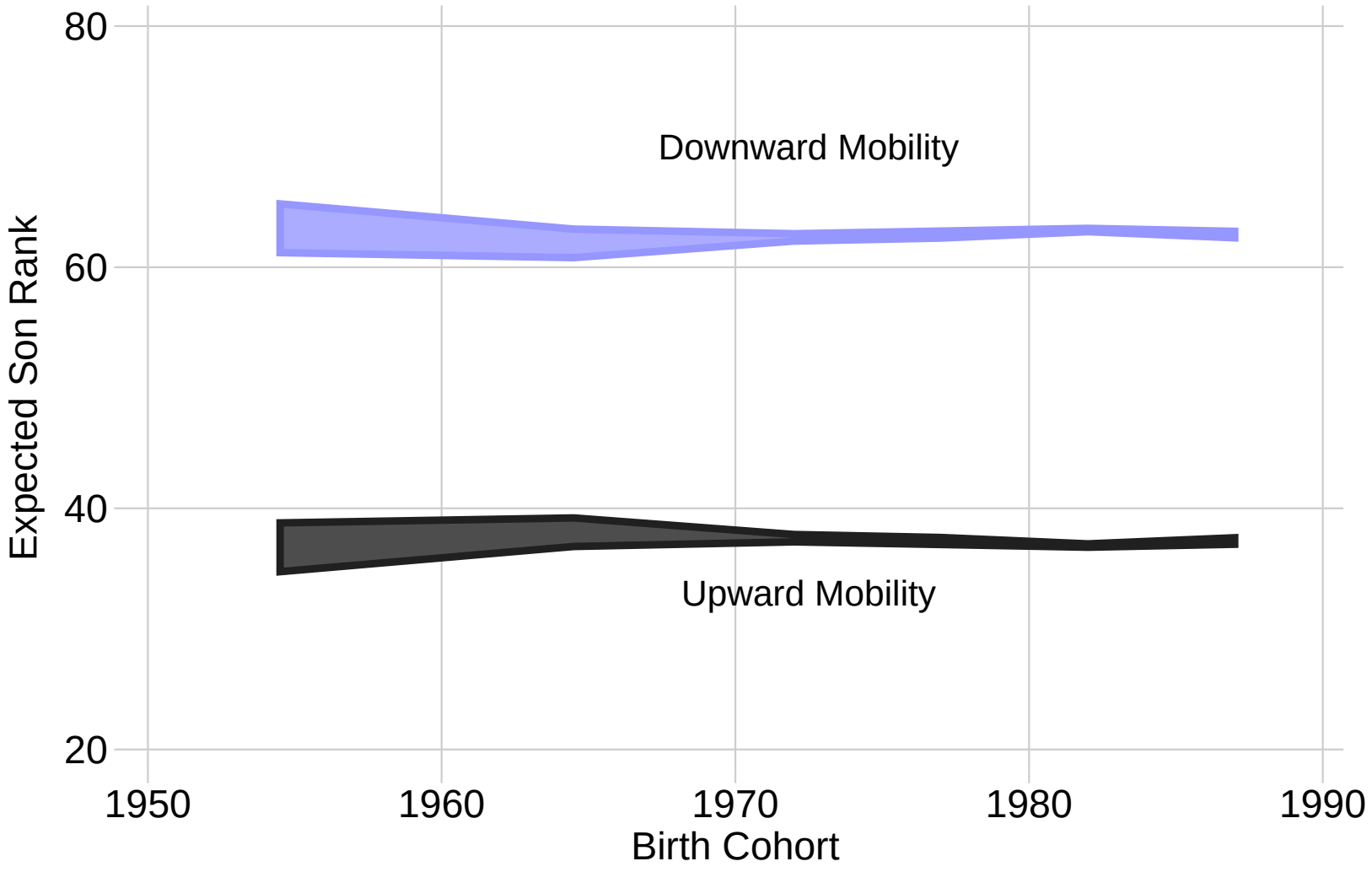


Comparing 1950s to 1980s birth cohort (naive estimate)

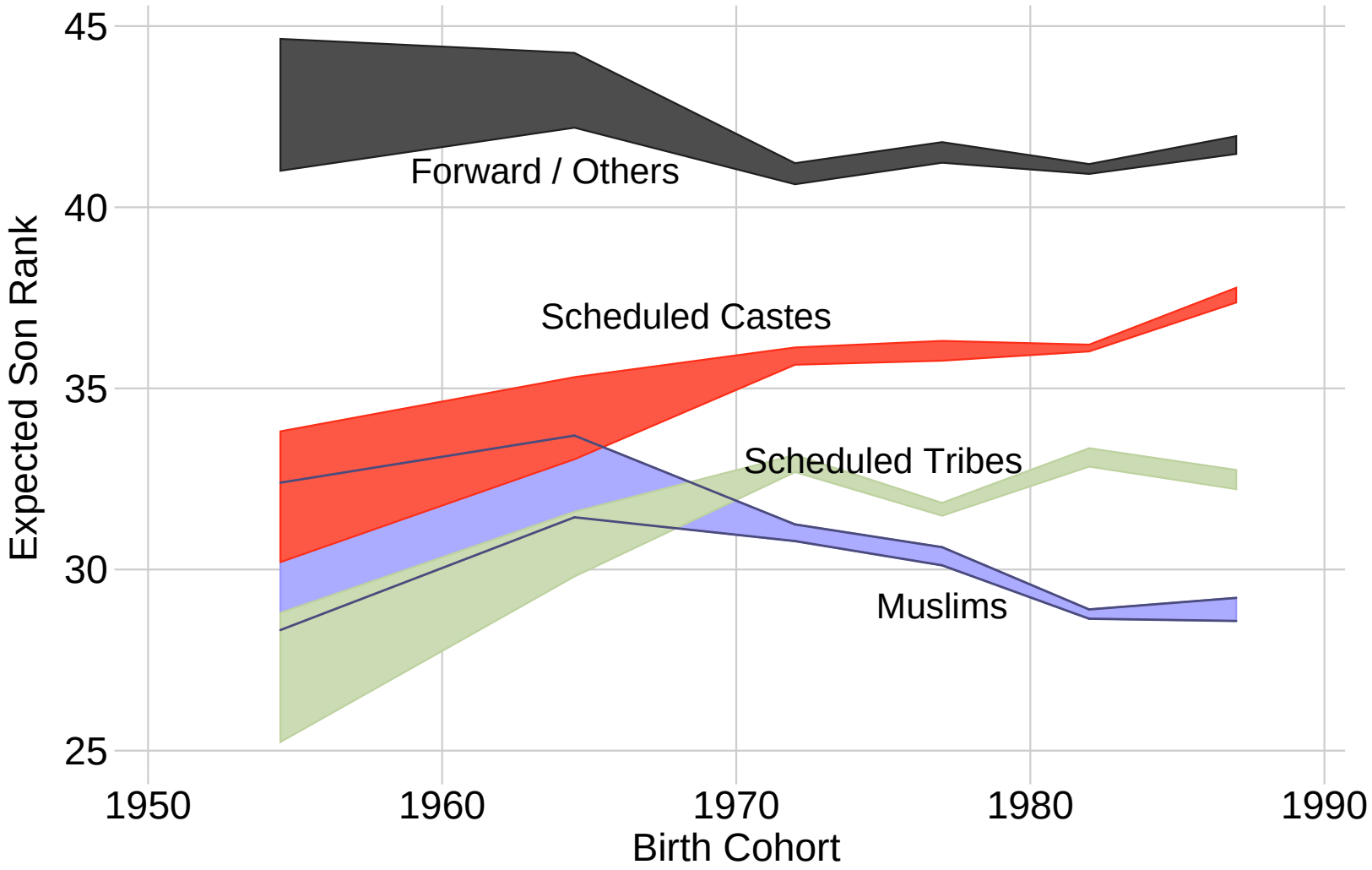


- Naive estimates suggests small but decisive mobility

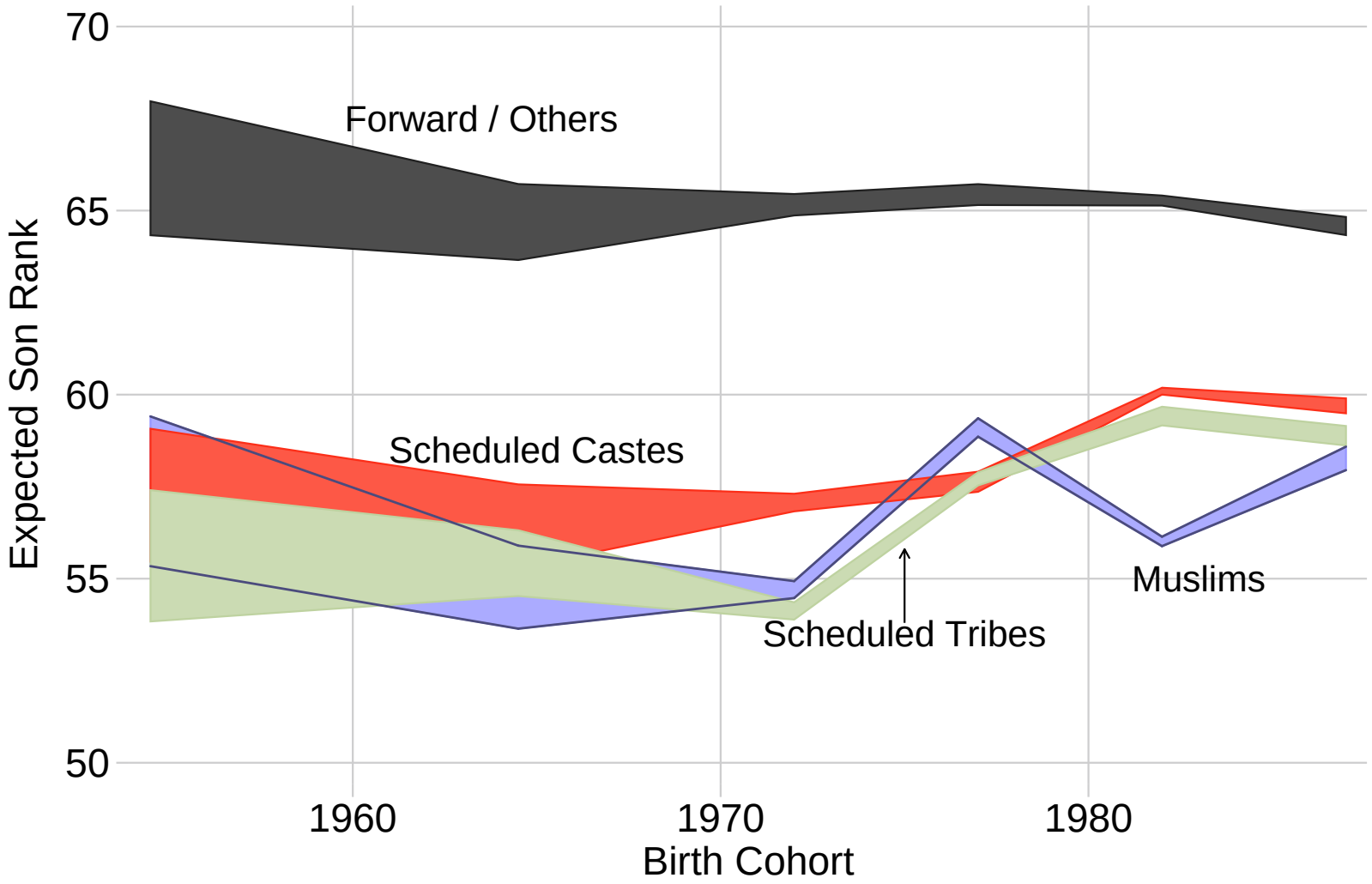
Upward and Downward Mobility over Time: All India



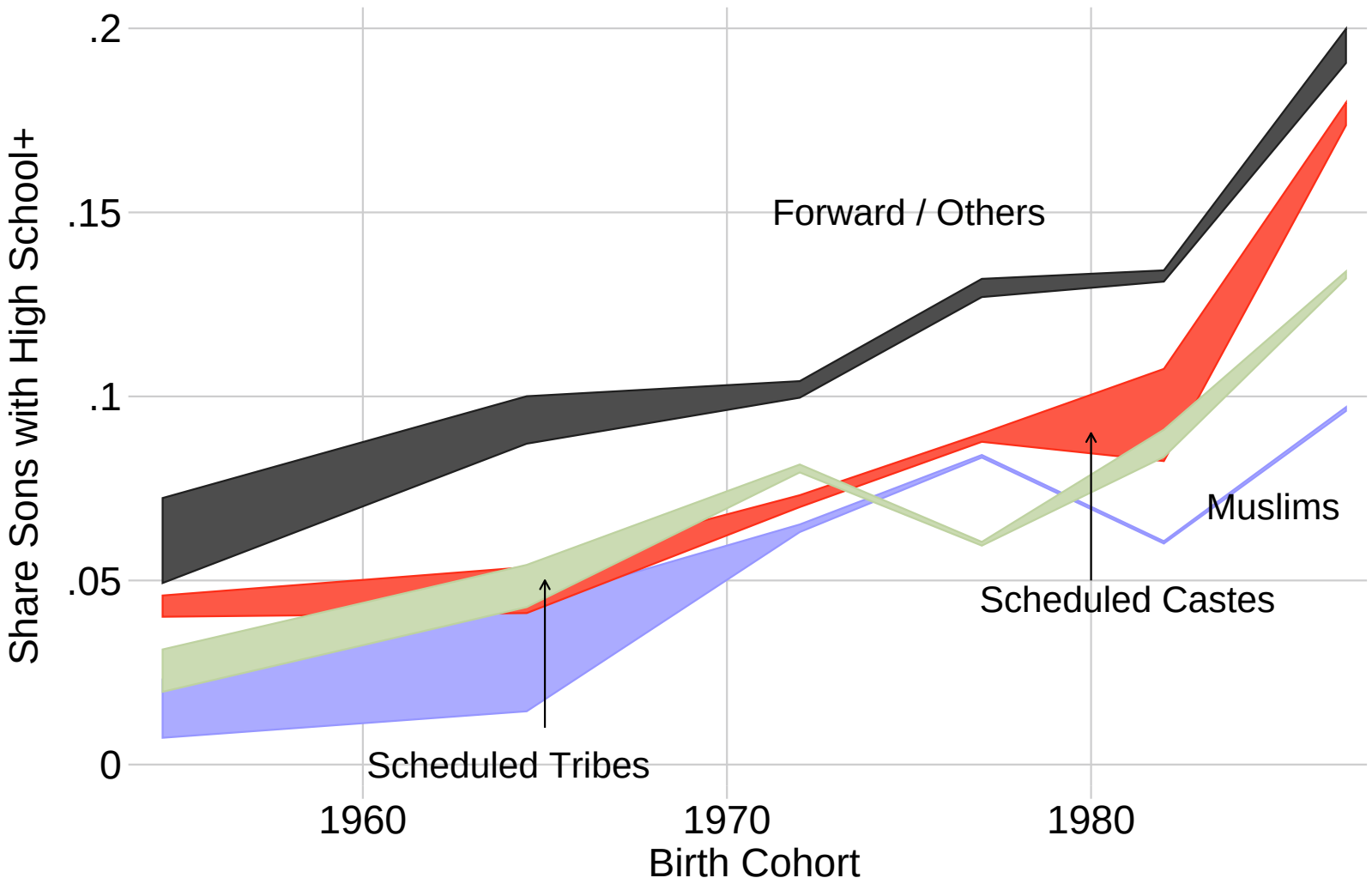
Upward Mobility: By Subgroup



Downward Mobility: By Subgroup



High School (μ_0^{50}, i.e. given Parent in Bottom Half)

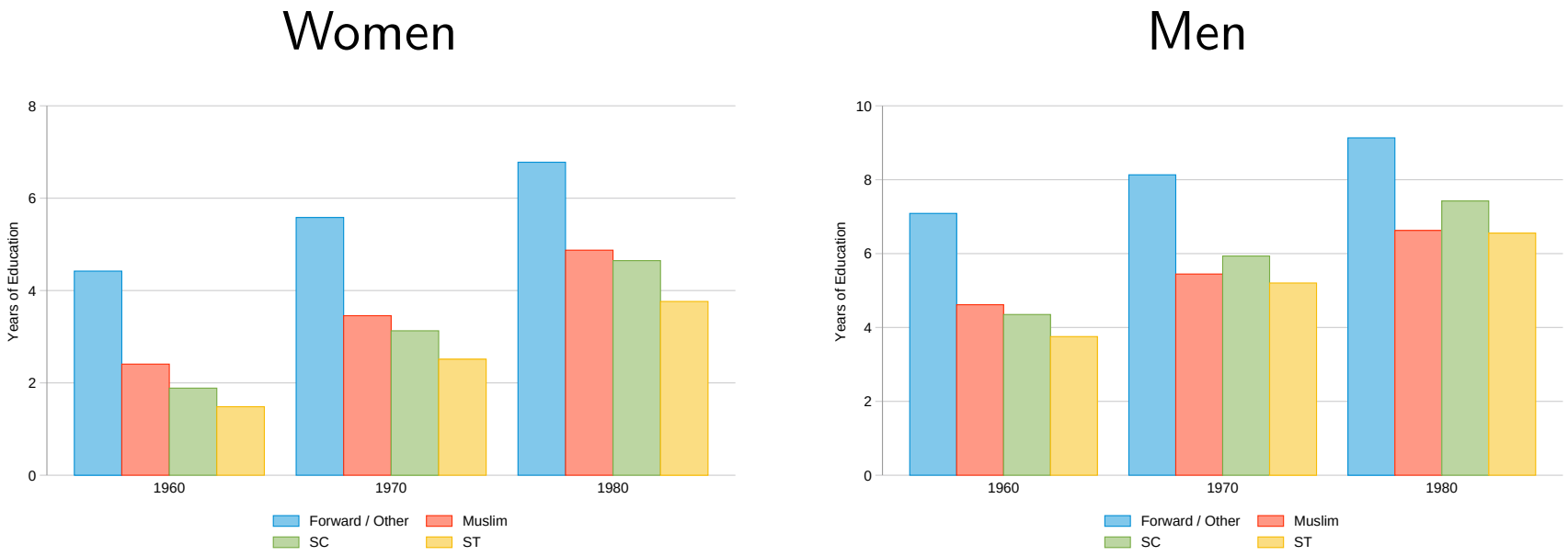


Primary School | Middle School | University

Gender and Mobility

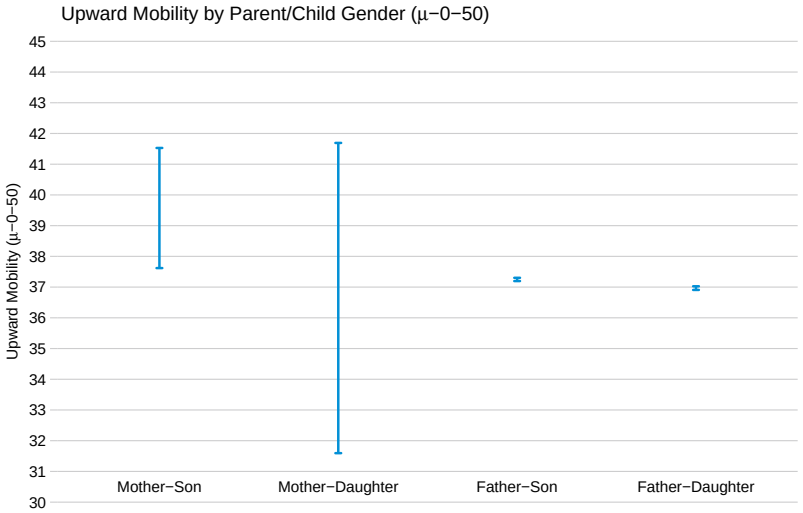
- Thus far all results for fathers and sons
- Breaking out results by gender tells a richer story:
  - Relative decline in Muslim education does not hold for women
  - Same-sex parents matter most for children's outcomes
  - Mother's education cannot explain decline in Muslim mobility

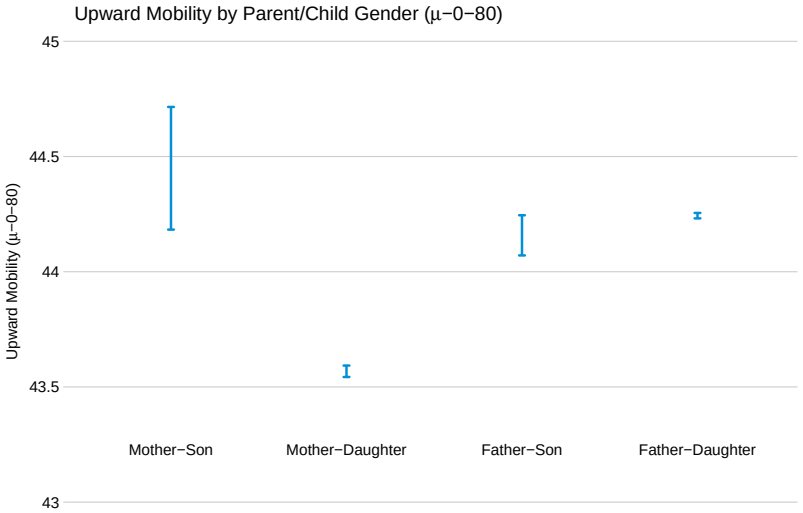
Educational Attainment by Group and Cohort (Full Sample)



- All groups gaining education
- Persistent gap between forward castes and others
- Muslim men but not women falling behind SCs

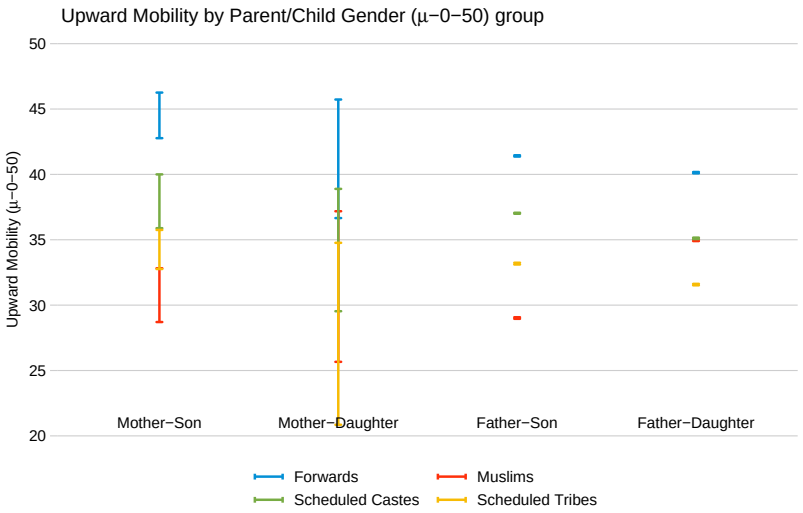
Mobility by Gender

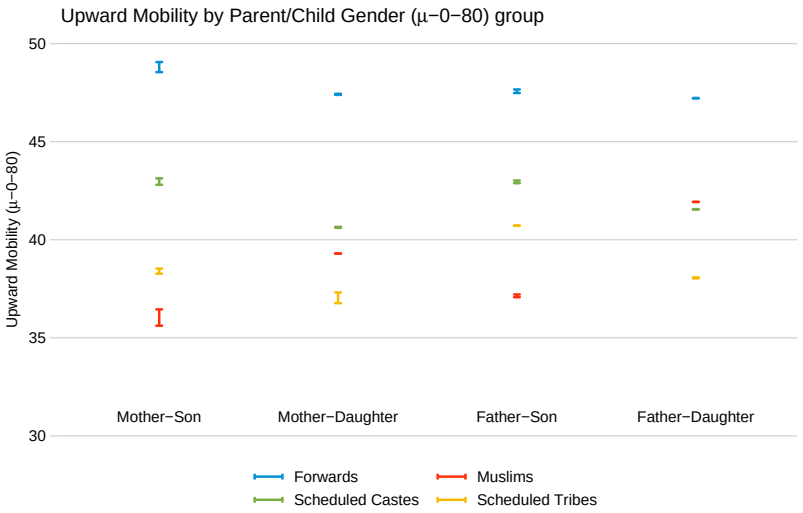




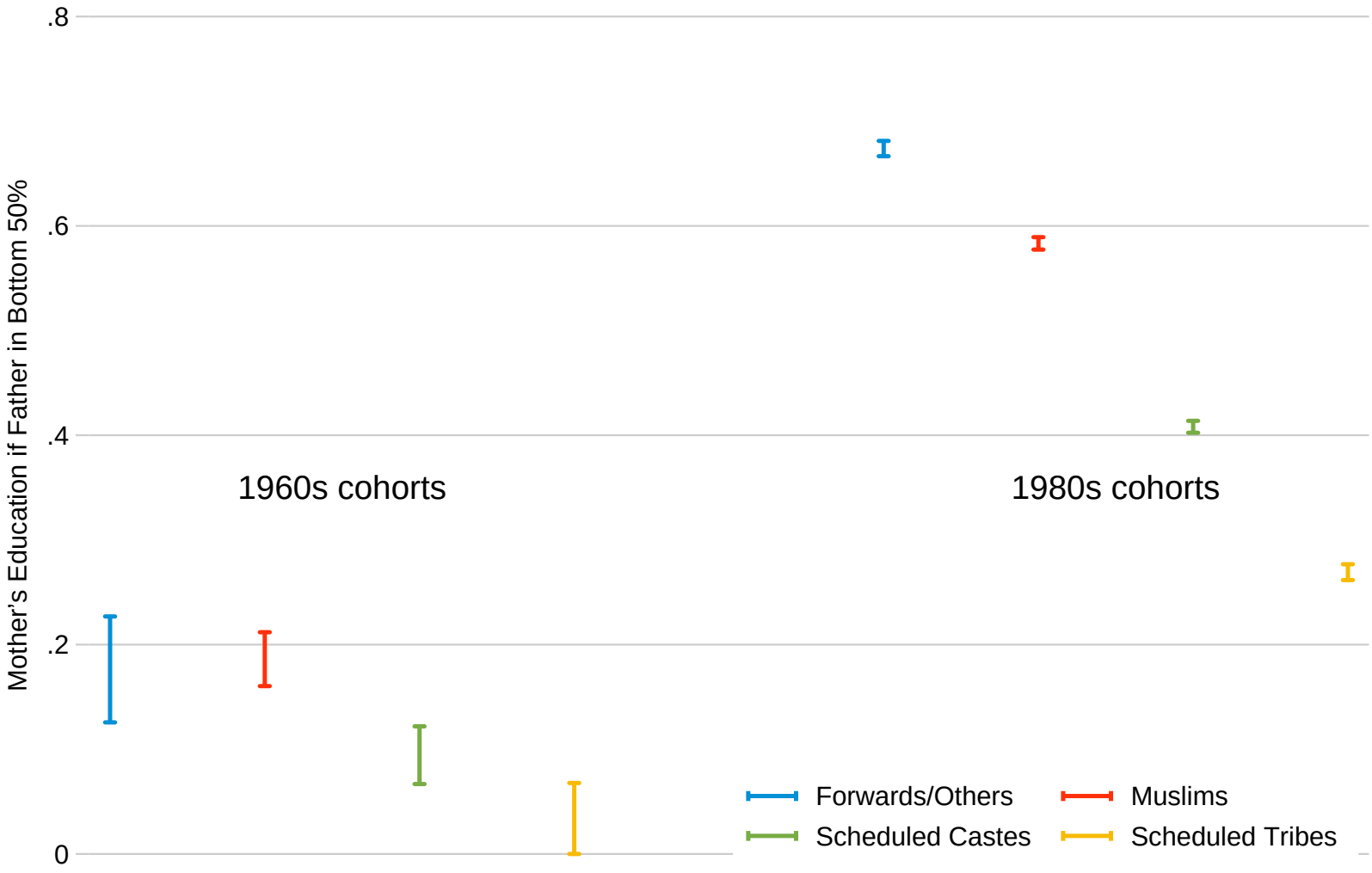
- Lower values mean that parent's low rank matters more for child outcomes
- Different intervals allow us to make different comparisons

Mobility by Gender and Group





Mother's Education Does Not Explain Muslim Decline



- Fertility differences also unable to explain the gap

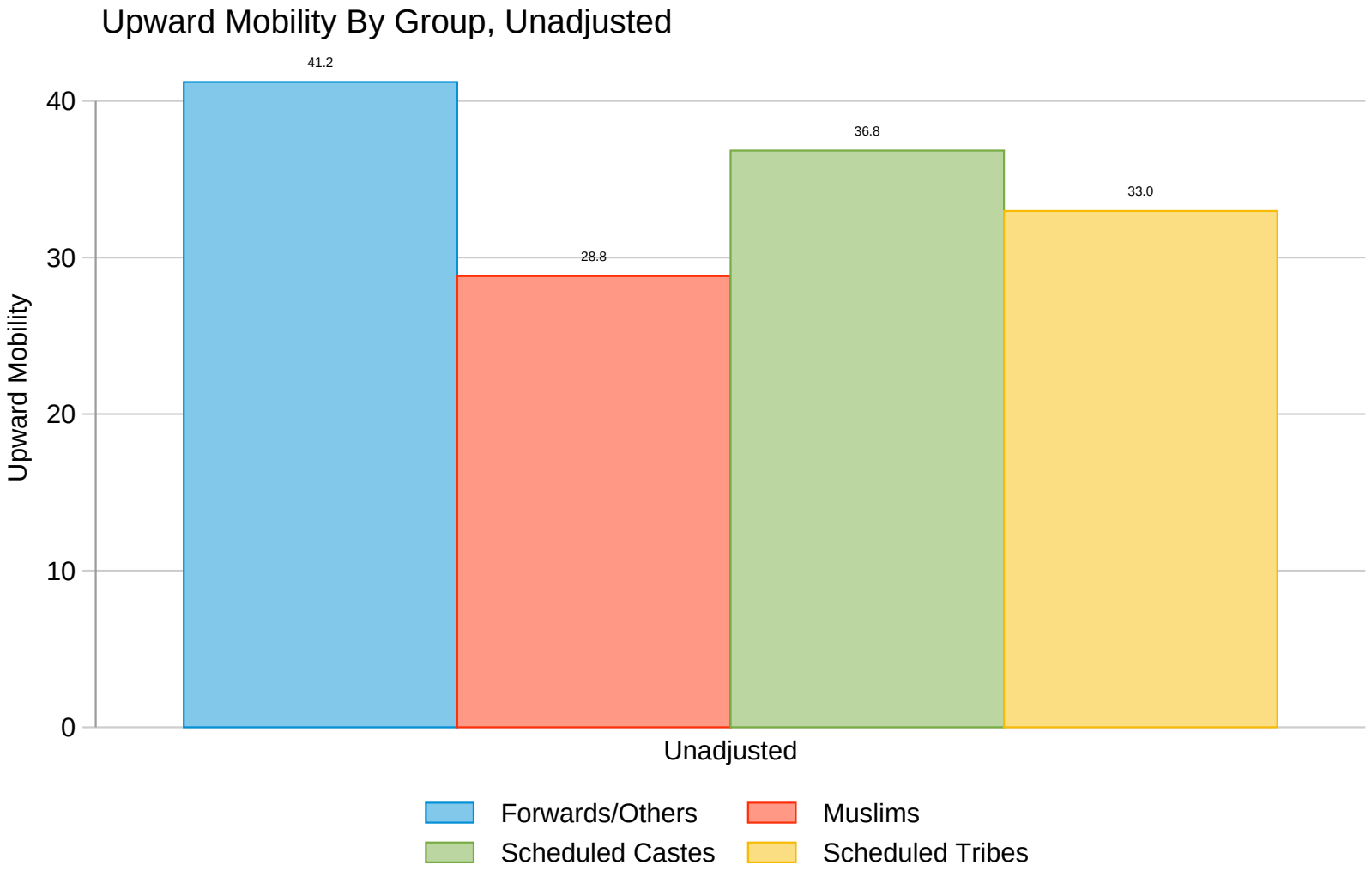
How Much Does Geography Explain?

Two extreme hypotheses:

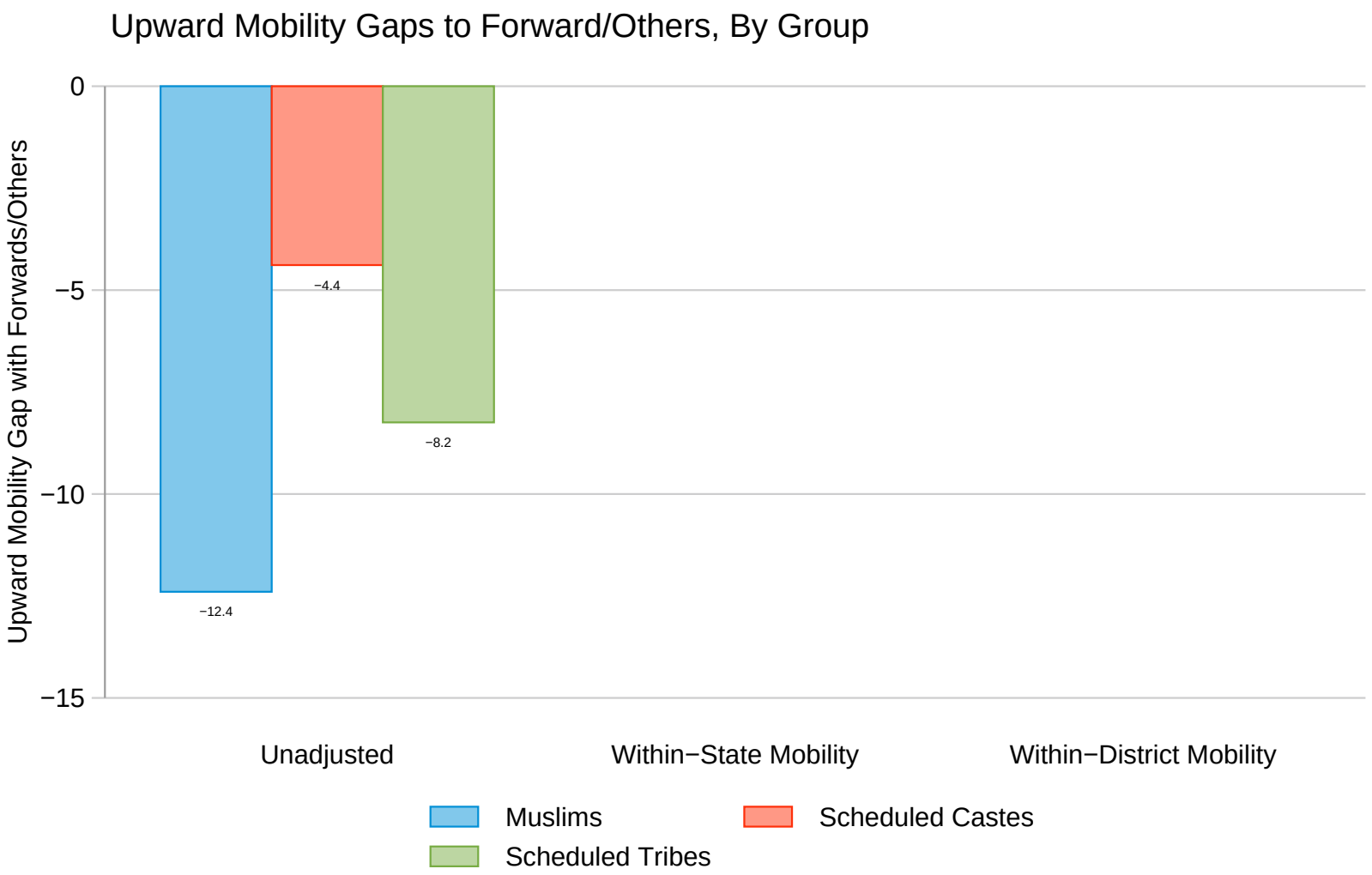
1. The mobility curve exists locally:
   - Within any location, boys with low-ranked fathers have little opportunity to advance
   - Within any location, mobility is highest for forwards and lowest for Muslims

2. Group differences are entirely based on geography:
   - *e.g.* Maybe everyone does well in AP, and poorly in Bihar
   - In places where Muslims live, everyone has low mobility
   - If we control for place fully, group differences will disappear

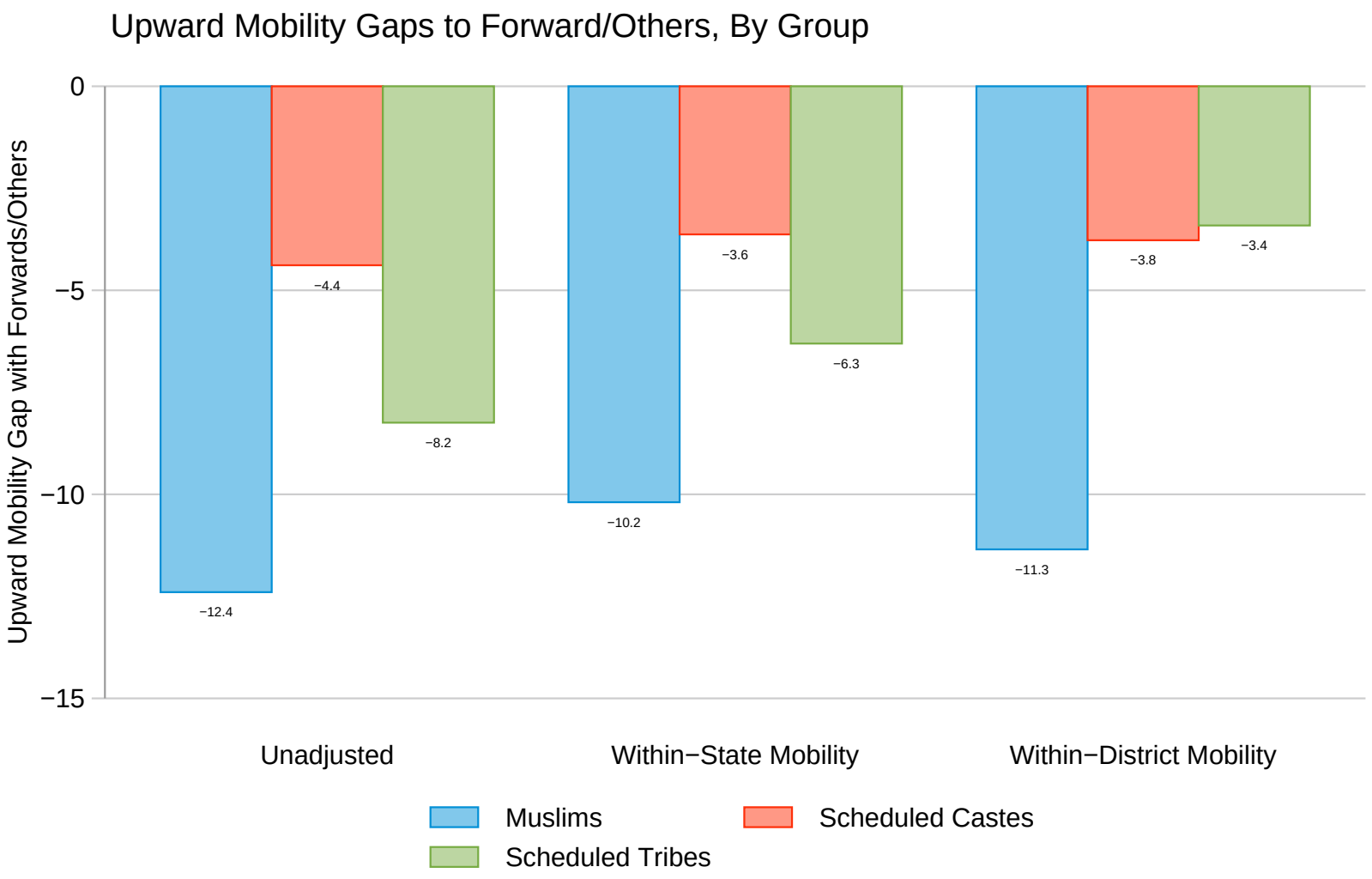
Mobility Differences Across Groups



Transform These to Mobility Gaps to Forward/Others



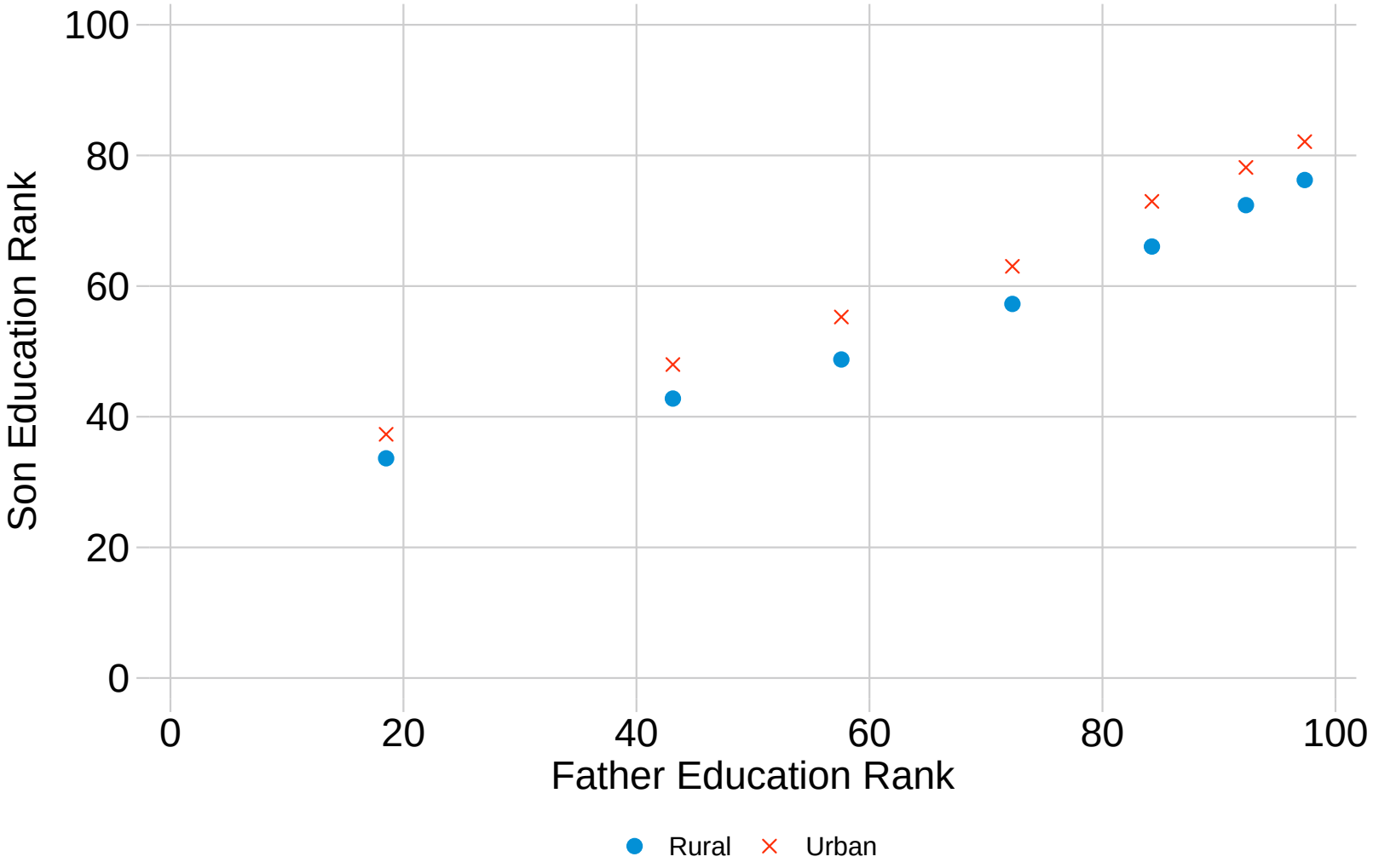
Mobility Gaps Controlling for State and District



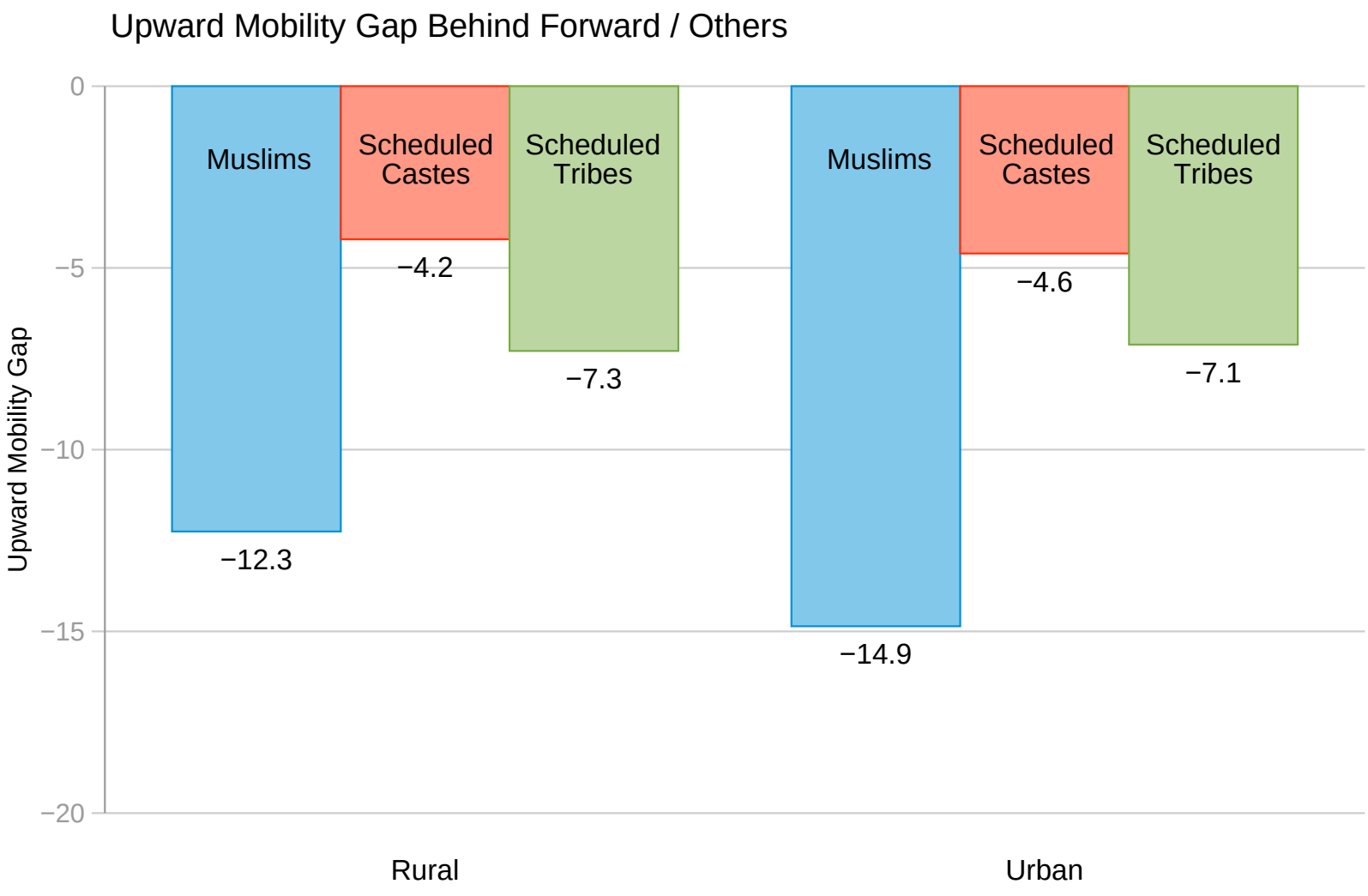
The Local Geography of Intergenerational Mobility

What characteristics of places predict high and low mobility?

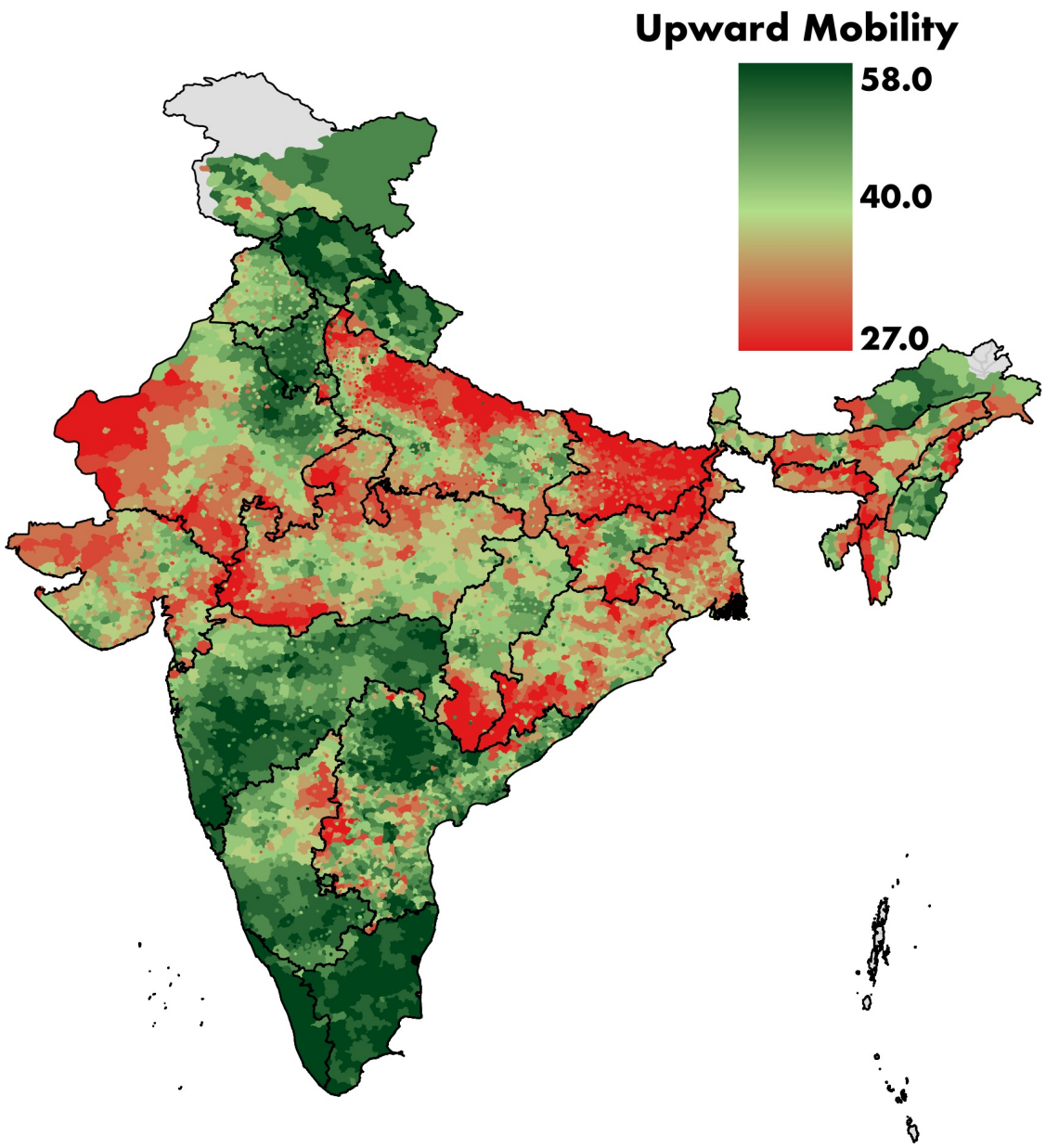
Mobility is Higher in Cities



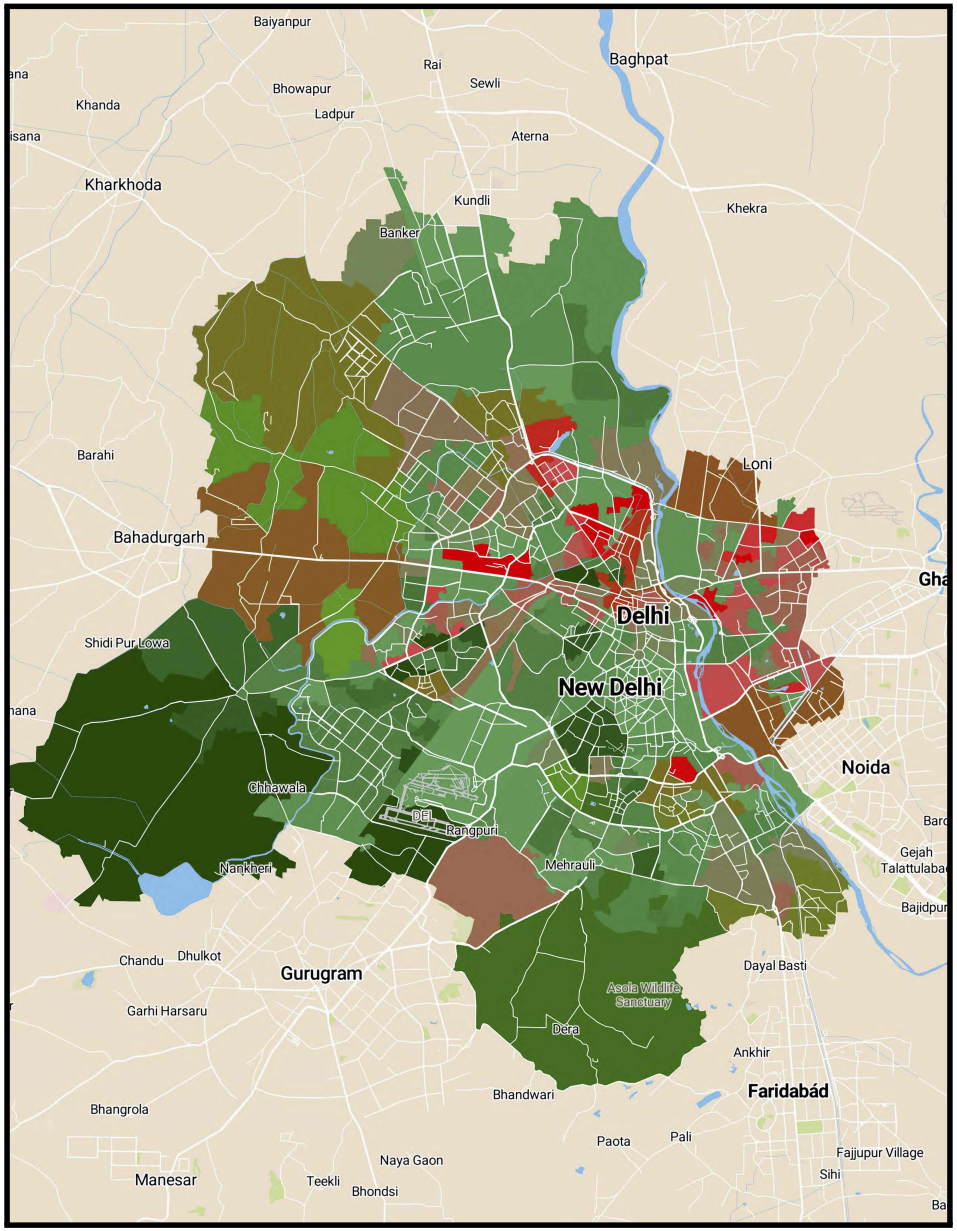
The Effect of Cities Depends on Group Identity



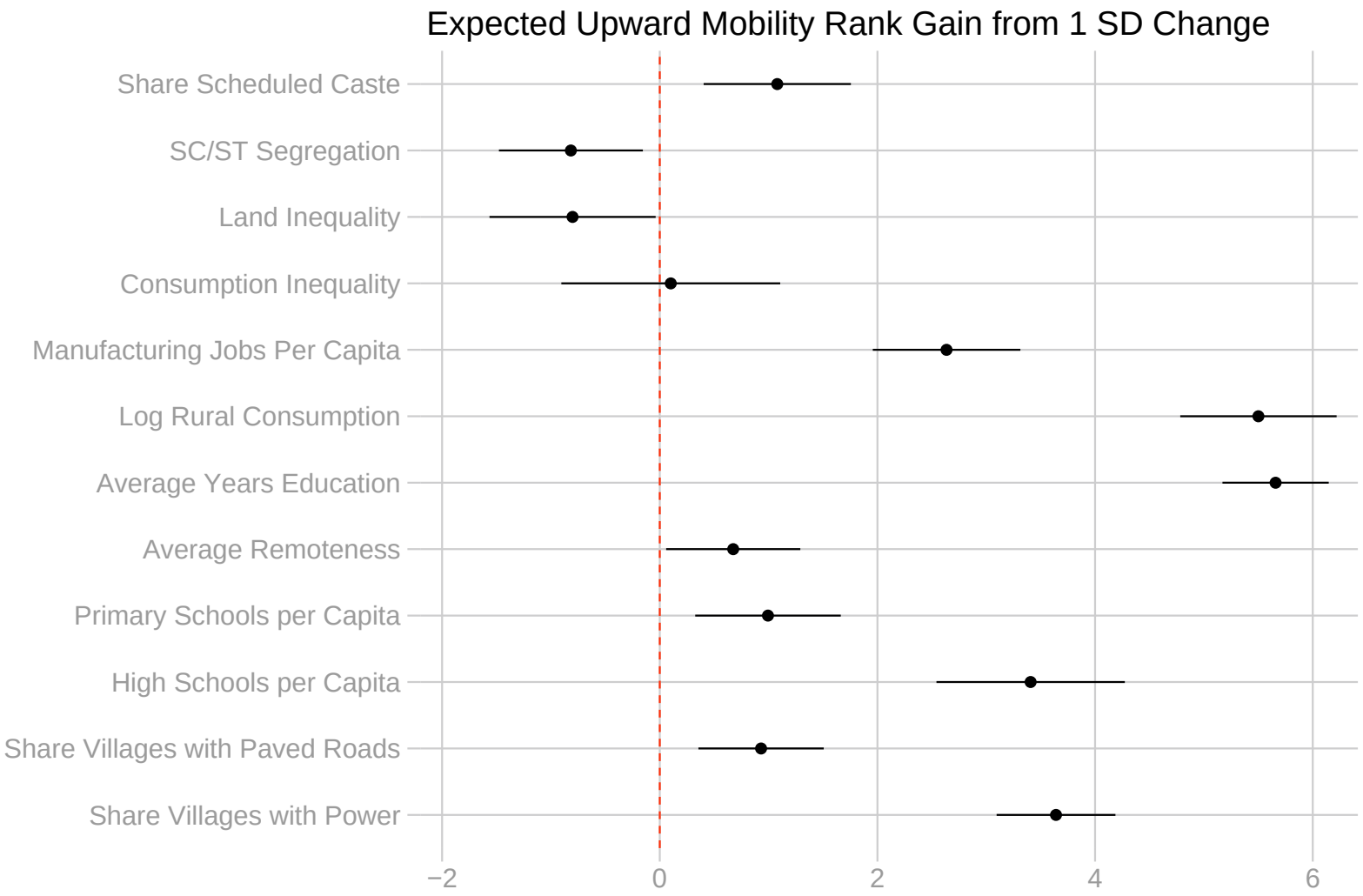
Regional Variation in Mobility is Substantial



Local Variation in Mobility is Also Substantial



Correlates of Rural Mobility



Correlates of Urban Mobility

Conclusions on Mobility in India

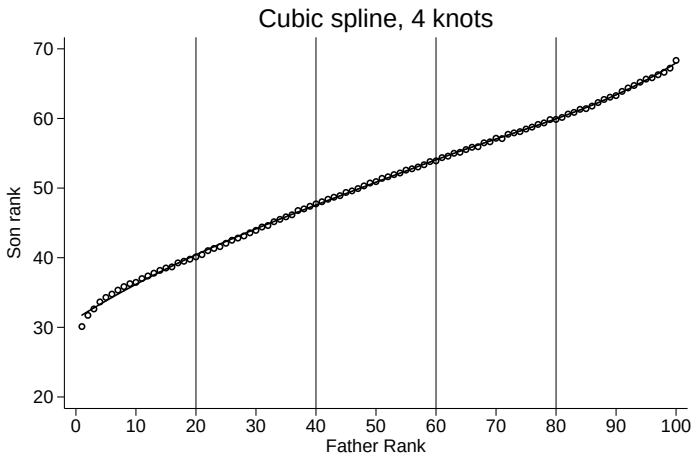
- Our method for calculating mobility is valid for comparison across subgroups, countries, and time when only educational data available for older generation
- Intergenerational mobility gains and growth are not inextricably linked
- Access to opportunity is improving for policy-targeted Scheduled Castes, but Muslims are losing ground
  - Contrasting political talking points on "Muslim appeasement"
- Substantial regional and local mobility variation

- Ongoing work on mechanisms: discrimination, affirmative action
- Long run: understanding barriers to migration, especially rural to urban

Appendix Figures

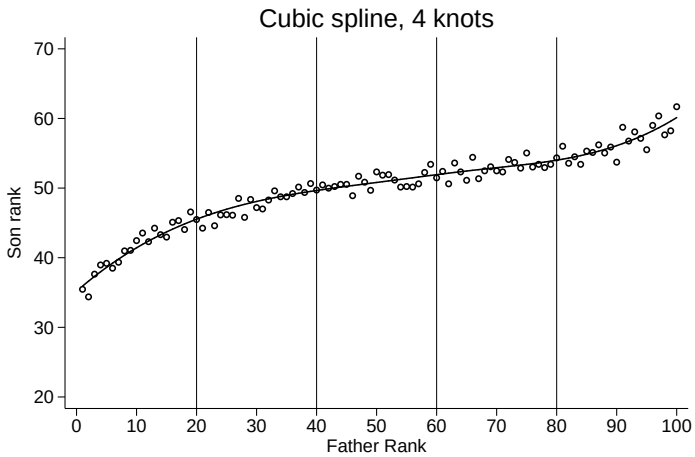
Choosing a Curvature Restriction

Panel A: U.S.A.



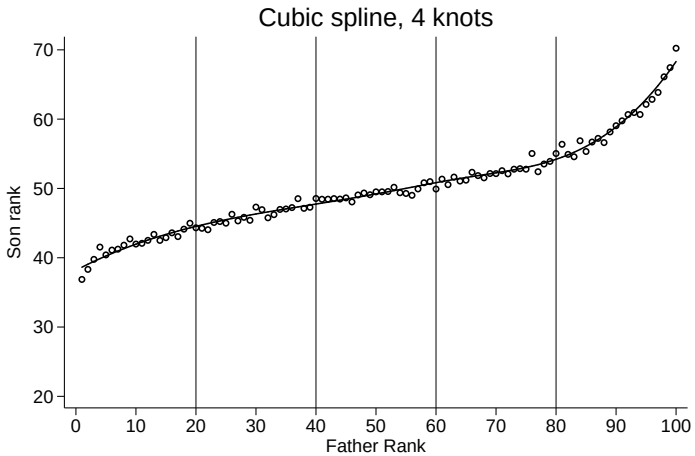
2nd derivative in [−.009,.017]

Panel B: Denmark



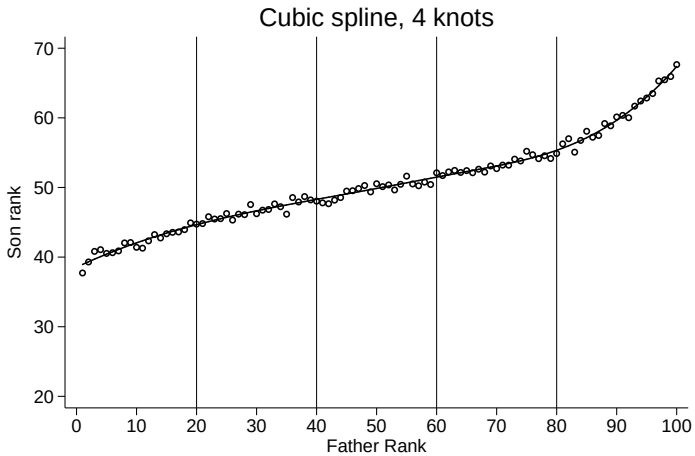
2nd derivative in [−.026,.028]

Panel C: Sweden



2nd derivative in [−.015,.066]

Panel D: Norway

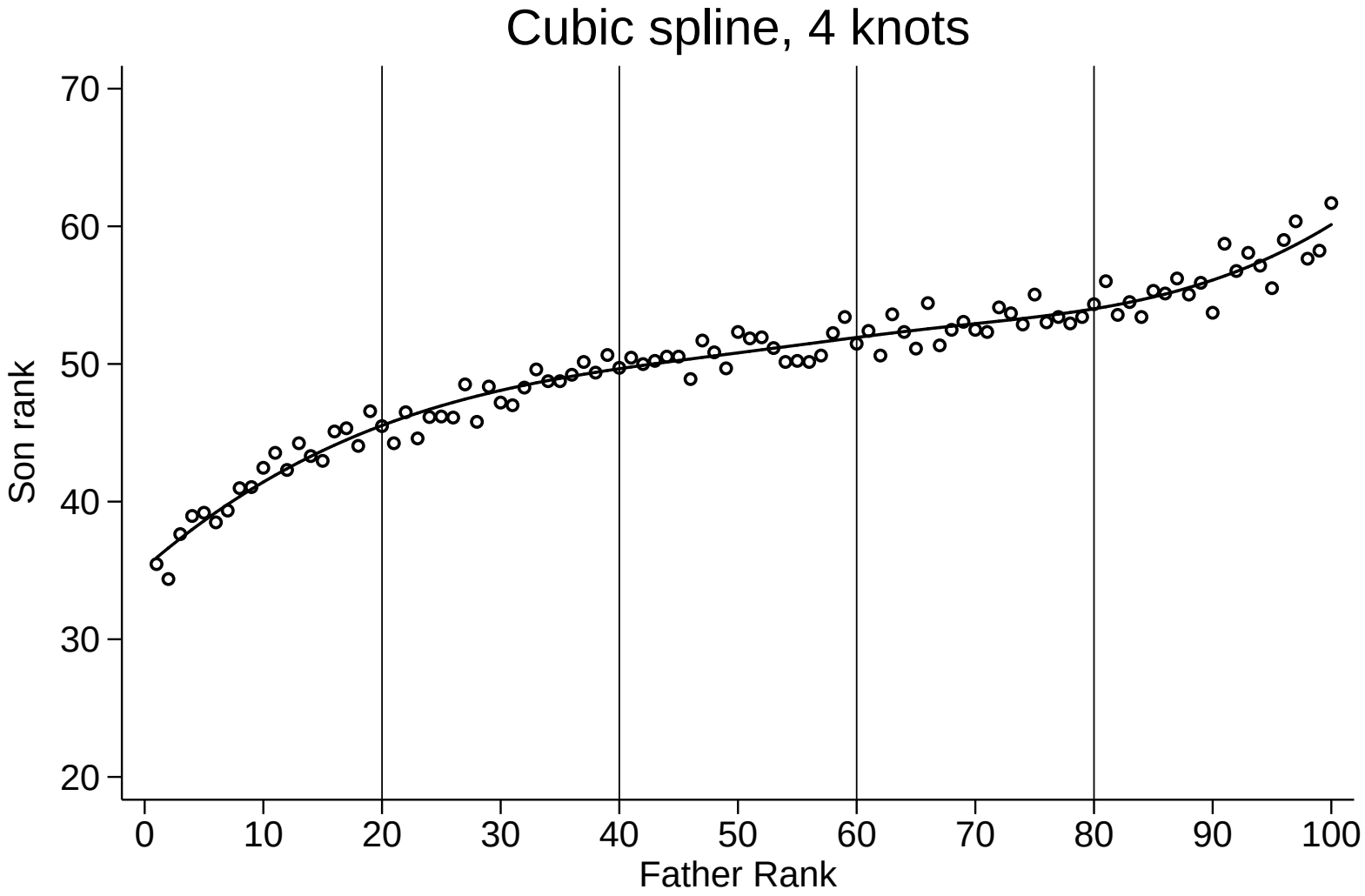


2nd derivative in [−.011,.046]

Test Case: Simulate Interval Censoring in Danish Data

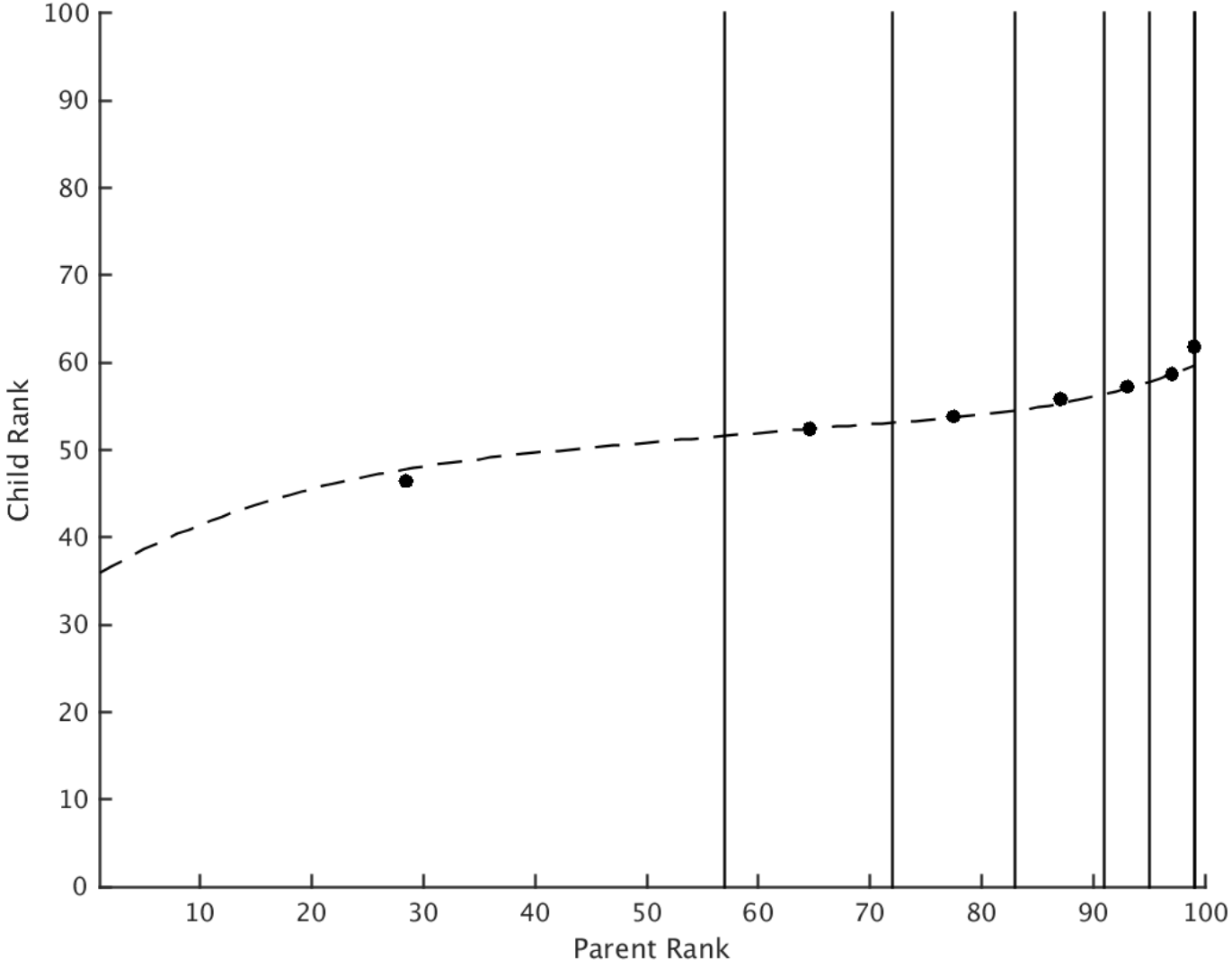
- We know the full CEF in Denmark
- Create parent bins on intervals from Indian data, and apply interval censoring
  - All we will observe is mean child rank in each parent bin
- Calculate bounds using our methodology and compare to original distribution

Test Case: Raw Danish Data

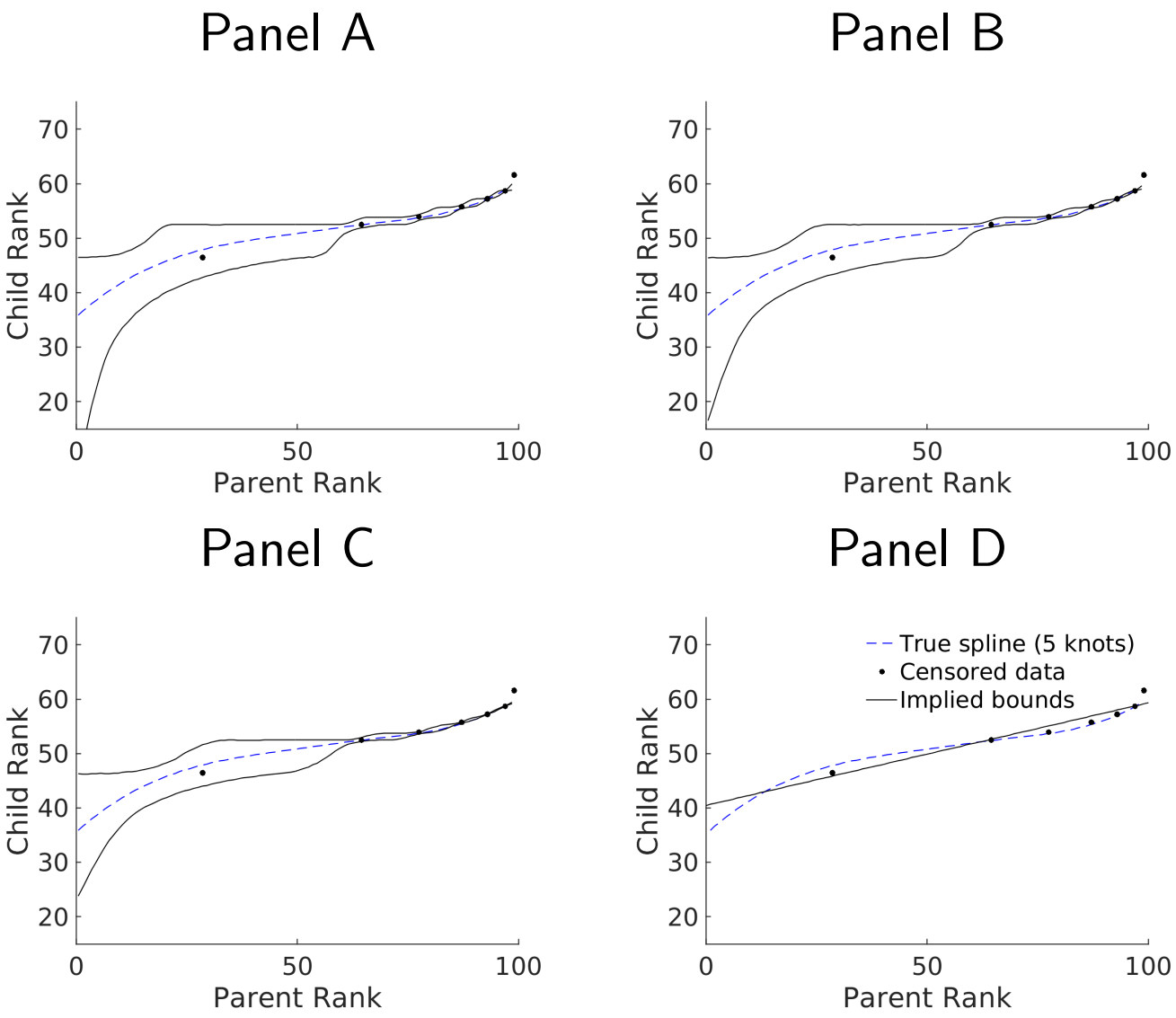


2nd derivative in [−.026,.028]

Test Case: Simulated Interval Censoring in Danish Data



Test Case: Recovering Bounds from Censored Data



Back

Implementing the Curvature Constraint

- Numerically calculate set of feasible curvature-constrained CEFs
- Minimize MSE with respect to data, i.e. match mean outcome in each coarse bin
- Calculate the min and maximum of the CEF at a given point that satifies this minimum MSE
- MSE $= 0$ is possible

Numerical Optimization Problem

We seek

$$Y^{min}(i) = \min\{Y(i)\} \qquad (1)$$

$$\text{such that}$$

$$Y(i) \text{ is weakly increasing in } i \qquad \text{(Monotonicity)}$$

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

$$\left[\sum_{k=1}^{K} \frac{i_{k+1} - i_k}{100} \left(\left(\frac{1}{i_{k+1} - i_k} \int_{i_k}^{i_{k+1}} Y(i) di \right) - r_k \right)^2 \right] = \underline{\text{MSE}}$$

(MSE Minimization)

- <u>MSE</u>: Lowest attainable *MSE* under monotonicity and curvature constraints
  - Could use any desired loss function

Back