Online Appendix A: Coresidency and Truncation Bias in Rural China: CFPS 2010 vs. CHIP 2002

An important advantage of the data used in this study is that the estimates are free of truncation bias that arises from using coresidency to define household membership in a survey. Emran, Greene, and Shilpi (2018) provide an in-depth discussion on the magnitude of truncation bias in the context of rural India. To the best of our knowledge, there are no studies on the effects of sample truncation due to coresidency in the context of rural China in the existing literature. To understand the extent of truncation bias in the context of rural China, we compare the estimates free of truncation in Tables 1 and 2 with the corresponding estimates using data from the Chinese Household Income Project (CHIP 2002). This may be of independent interest to many researchers as CHIP data sets have been widely used to study intergenerational mobility in China including by some of the authors of the current paper.

Using CHIP 2002 data, appendix Table A.2 reports estimates for two age ranges: (i) the same age cohorts as in Tables 1 and 2 in the main text of the paper (21-57 years old in 2002),[49] and (ii) the full sample with 18-65 years old children in 2002. The estimates of IGRC from CHIP 2002 are about 25 percent lower on average compared to the CFPS 2010 estimates (taking the CHIP 2002 estimate as the base). The estimates of the intercepts are, in contrast, biased upward in CHIP 2002, about 15-18 percent higher than the corresponding estimate from the CFPS data. The evidence that the coresident sample causes substantial downward bias in the slope estimates, but upward bias in the intercept estimates from the CHIP 2002 data is consistent with the evidence from India and Bangladesh provided by Emran, Greene and Shilpi (2018), and Emran and Shilpi (2018). The substantive conclusions are also less robust; the test of separability shows different results for IGRC and IRRS for the age cohorts overlapping with the main estimation samples from the CFPS used in Tables 1 and 2. While the null hypothesis of separability cannot be rejected for IGRC estimates at the 10 percent level, the IRRS estimates reject separability at the 5 percent level. An analysis based on the coresident sample thus can lead to wrong conclusions about the interaction between parent's education and occupation.[50]

[49]This ensures that the age cohorts in CHIP 2002 overlap with the 29-65 age cohorts used for CFPS 2010 data.

[50]A full analysis of CHIP 2002 data is presented in an unpublished work by Emran and Sun (2015b). As noted in the main text, the evidence and conclusions in this paper supersede those in Emran and Sun (2015b).

Online Appendix B: An Approach to Nature vs. Nurture

The approach developed here exploits the restrictions implied by the recent estimates of intergenerational correlation in cognitive ability in economics in a triangular biprobit model $a^{'}$ *la* Altonji et al. (2005), and performs sensitivity analysis over a narrow interval of the correlation parameter. This approach requires binary classification of educational attainment. We define a dummy D_i^p that takes on the value 1 when the father of a child has more than primary schooling and zero otherwise. Analogously we define a binary variable for children's education D_i^c which takes on the value of 1 when a child's schooling years is higher than a threshold (junior high schooling for China and secondary schooling for India). This means that we have a $[2 \times 2]$ occupation and education set-up. Consider the following triangular model of children's and parent's education:

$$Pr\left(D_i^c = 1\right) = \Phi\left(\lambda_0^j + \lambda_1^j D_i^p + X_i' \Pi^j + \varsigma_i\right)$$
$$Pr\left(D_i^p = 1\right) = \Phi\left(\gamma_0^j + X_i' \mathit{\Gamma}^j + \epsilon_i\right)$$

where Φ is the normal CDF and superscript j refers to parental occupation, i.e., $j \in \{f, n\}$. The error terms capture the unobserved factors including ability, part of which is genetically transmitted from father to son in our application.

The choice of the educational cut-offs described above is based on the following considerations. For parental generation, the cut-off is primary schooling, which equals 5 years of schooling in India and 6 years in China. For children in India, the cut-off we chose is 10 years of schooling that correspond to the completion of secondary schooling. The 10th year is a natural focal point for India as the students take one of the most important public examinations at the end of 10th grade. For China the cut-off used is junior high schooling which corresponds to 9 years of schooling. The 9th year cut-off is a focal point in China because of the 9-year compulsory schooling policy adopted by the government in 1986. While a lump in the distribution of children's schooling attainment at the 9th year of schooling may be driven by government policy, whether a child goes beyond 9 years of schooling would reflect more faithfully the effects of family background. Thus we define all the dummies as strict inequalities; for example, for rural China, we set $D^c = 1$ when a child attains *higher* than 9 years of schooling and zero otherwise. The conclusions below, however, do not depend on the precise cut-off chosen for the educational attainment dummies.[51]

[51] As part of robustness checks, we estimated biprobit models for China using 5 years cut-off for father

It is useful to decompose the error terms in the triangular model into two parts, one genetically transmitted (denoted as ϕ_i) and the other orthogonal to the genetic component:

$$\begin{aligned} \varsigma_i &= \phi_i^c + \tilde{\varsigma}_i \\ \epsilon_i &= \phi_i^p + \tilde{\epsilon}_i \end{aligned}$$

So $Cov\left(\varsigma_i, \epsilon_i\right) = Cov\left(\phi_i^c, \phi_i^p\right) + Cov\left(\tilde{\varsigma}_i, \tilde{\epsilon}_i\right)$. The focus is on the role played by the genetic component represented by $Cov\left(\phi_i^c, \phi_i^p\right)$.

We take advantage of a substantial literature in economics and behavioral genetics that provides estimates of correlation between parents and children in cognitive ability. According to the estimates reported by Black et al. (2009) and Bj rklund et al. (2010), $Cov\left(\phi_i^c, \phi_i^p\right) \in [0.35, 0.38]$. A meta-analysis of behavioral genetic studies by Plomin and Spinath (2004) reports a correlation of 0.40.

Let $\rho = Corr\left(\varsigma_i, \epsilon_i\right)$. We estimate the bivariate probit model above by imposing $\rho = 0.10, 0.20, 0.30, 0.40$.[52] Thus we allow for different degrees of correlation in the error terms of the triangular model driven by genetic transmissions across generations. It is important to appreciate that the estimates of intergenerational schooling persistence from the sensitivity analysis with $\rho = 0.40$ can be interpreted as the lower bounds on the true intergenerational effects due to economic and social factors. The estimates of ability correlations in the literature discussed above are likely to be biased upward as measures of *genetic correlations at birth* because they partly capture the dynamic interactions between family environment and genetics up to the age when the ability measurements are taken.[53] For example, the careful recent analysis by Gronqvist et al. (2017) uses measures of cognitive ability taken at age 13.

The results from the biprobit sensitivity analysis are reported in Table 3 of the main text and discussed at the beginning of section (6). We report the marginal effects of

and 10 years cut-off for children, so that the educational attainment dummies in China match exactly those in India in terms of years of schooling. The estimates are very close to those reported in Table 3 of the main text. The details are available from the authors.

[52]In an interesting recent paper, Gronqvist et al. (2017) show that the estimates of ability correlation may be biased downward because of measurement error, and they report estimates in the range of 0.42-0.48. However, note that the educational attainment as measured by years of schooling is also usually measured with error. It thus seems reasonable to use 0.40 as the upper bound for our analysis, assuming that the measurement errors in schooling and cognitive ability largely offset each other.

[53]This includes the in-utero effects on a child's health which is found to be of substantial magnitude according to a large literature. Please see the survey by Almond and Currie (2011). Recent evidence shows that socioeconomic status a child born into affects the development of brain in a significant way (Noble et al. (2015)).

switching the parental educational dummy from zero (primary or below) to one (more than primary) which implies that the estimates of intergenerational persistence depend on all of the probit coefficients in the model, including the intercept term.[54] The main text of the paper provides a concise discussion of the main results from this analysis. Here we provide a more detailed discussion.

Effects of Positive Ability Correlations in Rural India

The estimates from the univariate probit model for India reported in row 1 of Table 3 of the main text shows that the pattern of the effects is broadly similar to that found in the estimates based on years of schooling in Table 1 and Table 2 in the main text. The sons born to fathers with more than primary schooling enjoy a higher probability of attaining more than secondary schooling, and this effect is larger for the children with fathers in nonfarm occupation.

More important and interesting for our purpose is how the estimates are affected when we relax the assumption of $\rho = 0$ imposed in the univariate probit estimates. Consistent with *a priori* expectation, the magnitude of the estimated effects declines monotonically with higher values of intergenerational ability correlation ρ. However, the estimated effect of having a better educated father remains positive and statistically significant at the 1 percent level across the board even when we set the value of ρ at its upper bound of 0.40. This can be interpreted as strong evidence that the estimated effects in India are not likely to be driven exclusively by mechanical effects of genetic transmission of ability.

Effects of Positive Ability Correlations in Rural China

The estimates of father's effect in rural China for alternative values of genetic correlation in ability are reported in the first two columns of Table 3 in the main text. The estimates from the univariate probit model under the assumption that $\rho = 0$ show that the effects for both farm and nonfarm households in China are much smaller in magnitude when compared to the corresponding estimates for India. This confirms the conclusion based on years of schooling as a measure of educational attainment that the sons in rural China enjoy much higher educational mobility compared to the sons in rural India. More important is the evidence that the estimated effects become very small in China when we set $\rho = 0.20$. The advantage due to the higher educated father declines effectively to zero when we set $\rho = 0.30$, as the estimated effects turn negative for both the farm and nonfarm households.

[54] The AET sensitivity analysis including the quadratic age controls yield similar conclusions to those based on Table 3 in the main text. The results are available from the authors upon request.

The evidence thus suggests that the estimates of intergenerational persistence in China reported in Table 1 and Table 2 in the main text could be entirely driven by genetic correlations in ability. This suggests high intergenerational educational mobility during the decades of 1980s and 1990s in rural China, for the sons of both the farmer and non-farmer fathers. However, as discussed in details in the main text (see sections (6.1) and (6.2)), the evidence that the observed persistence in rural China could be explained away by genetic correlation in cognitive ability does not necessarily imply that economic forces did not play a role in determining the pattern across farm and nonfarm households.

APPENDIX TABLES

Table A.1: Summary Statistics
(Main Overlapping Samples)

	Mean	SD	Min	Max
		CFPS 2010		
Son's Schooling (years)	**6.31**	**4.14**	**0**	**19**
Father's Schooling (years)	**3.81**	**4.14**	**0**	**16**
Father in Agriculture (dummy)	**0.48**	**0.50**	**0**	**1**
Son's Age (years)	**39.78**	**7.69**	**29**	**65**
Father's Age (years)	**67.92**	**8.83**	**45**	**96**
		REDS 1999		
Son's Schooling (years)	**6.26**	**5.14**	**0**	**21**
Father's Schooling (years)	**4.13**	**4.41**	**0**	**20**
Father in Agriculture (dummy)	**0.59**	**0.49**	**0**	**1**
Son's Age (years)	**28.65**	**7.88**	**18**	**54**
Father's Age (years)	**59.67**	**9.86**	**34**	**98**

NOTES: (1) CFPS stands for China Family Panel Studies and REDS stand for Rural Economic and Demographic Survey. (2) SD stands for standard deviation. (3) No. of observations for CFPS is 3305, and for REDS 6887.

Table A.2: Estimates of Intergenerational Educational Persistence in Rural China from CHIP 2002

	Panel A: Children 18-65 Years			
	SLOPE		INTERCEPT	
	IGRC (ψ_1^j)	IRRS (δ_1^j)	(ψ_0^j)	(δ_0^j)
Combined	**0.265**	**0.295**	**7.105**	**0.357**
	(0.018)	(0.019)	(0.163)	(0.016)
Farm	**0.245**	**0.263**	**7.075**	**0.353**
	(0.025)	(0.026)	(0.202)	(0.019)
Non-farm	**0.266**	**0.305**	**7.29**	**0.376**
	(0.024)	(0.027)	(0.219)	(0.022)
	Test of Separability		Test of Equality	
	($\psi_1^n = \psi_1^f$)	($\delta_1^n = \delta_1^f$)	($\psi_0^n = \psi_0^f$)	($\delta_0^n = \delta_0^f$)
F-statistics	0.410	1.375	0.700	0.917
P-value	0.521	0.243	0.403	0.340

	Panel B: Children of 21-57 years age in 2002			
	SLOPE		INTERCEPT	
	IGRC (ψ_1^j)	IRRS (δ_1^j)	(ψ_0^j)	(δ_0^j)
Combined	**0.262**	**0.283**	**7.074**	**0.360**
	(0.019)	(0.021)	(0.174)	(0.017)
Farm	**0.226**	**0.237**	**7.127**	**0.361**
	(0.027)	(0.027)	(0.216)	(0.021)
Non-farm	**0.277**	**0.310**	**7.180**	**0.372**
	(0.026)	(0.028)	(0.234)	(0.023)
	Test of Separability		Test of Equality	
	($\psi_1^n = \psi_1^f$)		($\psi_0^n = \psi_0^f$)	
F-statistics	2.216	4.205	0.039	0.189
P-value	0.139	0.043	0.844	0.664

Table A.3: Evidence on Complementarity and Intercept Differences with Quadratic Age Controls (Years of Schooling)

	Estimates with Son's Age and Age Squared as Controls			
	China		India	
	Farm	Nonfarm	Farm	Nonfarm
IGRC (ψ_1^j)	0.306	0.308	0.489	0.556
H_0 : Farm = Nonfarm ($\psi_1^n = \psi_1^f$)				
F Statistic	0.004		3.785	
P-value	0.950		0.053	
Intercept (ψ_0^j)	7.502	7.885	5.146	4.515
H_0 : Farm = Nonfarm ($\psi_0^n = \psi_0^f$)				
F Statistic	1.790		6.053	
P-value	0.183		0.015	

	Estimates with Both Son's and Father's Age and Age Squared			
	China		India	
	Farm	Nonfarm	Farm	Nonfarm
IGRC (ψ_1^j)	0.301	0.314	0.497	0.569
H_0 : Farm = Nonfarm ($\psi_1^n = \psi_1^f$)				
F Statistic	0.141		4.330	
P-value	0.708		0.039	
Intercept (ψ_0^j)	14.43	14.80	4.329	3.536
H_0 : Farm = Nonfarm ($\psi_0^n = \psi_0^f$)				
F Statistic	1.677		9.308	
P-value	0.198		0.025	

Table A.4: Evidence on Complementarity and Intercept Differences with Quadratic Age Controls (Schooling Rank)

	Estimates with Son's Age and Age Squared as Controls			
	China		India	
	Farm	Nonfarm	Farm	Nonfarm
IRRS (δ_1^j)	0.328	0.341	0.421	0.499
H_0 : Farm = Nonfarm ($\delta_1^n = \delta_1^f$)				
F Statistic	0.119		6.377	
P-value	0.731		0.012	
Intercept (δ_0^j)	0.456	0.479	0.370	0.314
H_0 : Farm = Nonfarm ($\delta_0^n = \delta_0^f$)				
F Statistic	0.761		7.433	
P-value	0.385		0.007	

	Estimates with Both Son's and Father's Age and Age Squared			
	China		India	
	Farm	Nonfarm	Farm	Nonfarm
IRRS (δ_1^j)	0.322	0.350	0.428	0.511
H_0 : Farm=Nonfarm ($\delta_1^n = \delta_1^f$)				
F Statistic	0.578		7.199	
P-value	0.449		0.008	
Intercept (δ_0^j)	0.928	0.946	0.303	0.237
H_0 : Farm=Nonfarm ($\delta_0^n = \delta_0^f$)				
F Statistic	0.456		10.27	
P-value	0.501		0.002	

Table A.5: Father's Education and Household Income (Control for Number of Children)

Panel A: Estimates for Rural China

	Intercept		Slope	
	Farm	**Nonfarm**	**Farm**	**Nonfarm**
CHIP 2002	**5654.30**	**8311.39**	**294.37**	**339.45**
	(747.55)	(1418.46)	(89.78)	(114.91)
CHIP 1995	**3434.95**	**3242.18**	**43.66**	**114.72**
	(523.84)	(617.94)	(68.88)	(44.65)

	Test of Equality Between Farm and Nonfarm	
	H_0: Intercepts are Equal	**H_0: Slopes are Equal**
CHIP 2002		
F Statistic	**0.16**	**0.03**
P-value	**0.69**	**0.87**
CHIP 1995		
F Statistic	**0.06**	**0.94**
P-value	**0.81**	**0.34**

Panel B: Estimates for Rural India

	Intercept		Slope	
	Farm	**Nonfarm**	**Farm**	**Nonfarm**
NSS 1993	**823.87**	**770.09**	**63.46**	**75.60**
	(11.49)	(16.16)	(2.48)	(3.10)

	Test of Equality Between Farm and Nonfarm	
	H_0: Intercepts are Equal	**H_0: Slopes are Equal**
F Statistic	**591.19**	**9.98**
P-value	**0.000**	**0.002**

Notes: (1) The dependent variable for Rural China is the average household income (total). CHIP 2002 is the average of the last 5 years of total household income, and CHIP 1995 is the average of the last 3 years of household income. The dependent variable for India is total household expenditure. (2) The numbers in parenthesis are standard errors. (3) H_0 stands for Null Hypothesis. (4) The number of observations for CHIP 1995: Farm (1709), Nonfarm (3893), and for CHIP 2002: Farm (4457), Nonfarm (4087). For NSS (1993), the number of observations is Nonfarm (20535), and Farm (48196).