WPS7608

Policy Research Working Paper 7608

When Measure Matters

Coresidency, Truncation Bias, and Intergenerational Mobility in Developing Countries

M. Shahe Emran
William Greene
Forhad Shilpi

WORLD BANK GROUP
Development Research Group
Environment and Energy Team
March 2016

Policy Research Working Paper 7608

Abstract

Biases from truncation caused by coresidency restriction have been a challenge for research on intergenerational mobility. Estimates of intergenerational schooling persistence from two data sets show that the intergenerational regression coefficient, the most widely used measure, is severely biased downward in coresident samples. But the bias in intergenerational correlation is much smaller, and is less sensitive to the coresidency rate. The paper provides explanations for these results. Comparison of intergenerational mobility based on the intergenerational regression coefficient across countries, gender, and over time can be misleading. Much progress on intergenerational mobility in developing countries can be made with the available data by focusing on intergenerational correlation.

This paper is a product of the Environment and Energy Team, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at fshilpi@worldbank.org.

When Measure Matters: Coresidency, Truncation Bias, and Intergenerational Mobility in Developing Countries[1]

M. Shahe Emran

IPD, Columbia University

William Greene

New York University

Forhad Shilpi

DECRG, World Bank

Key Words: Coresidency, Truncation Bias, Intergenerational Mobility, Developing Countries, Intergenerational Regression Coefficient (IGRC), Intergenerational Correlation (IGC), LSMS, HIES, Bangladesh, India

JEL Codes: O12, J62

---

[1] We would like to thank Matthew Lindquist, Hector Moreno, Tom Hertz, Claudia Berg and seminar participants at NEUDC 2015 at Brown University for helpful comments on earlier versions, and Gabriela Aparicio for help with data at the early stage of this project. An earlier version of the paper was circulated under the title "When Measure Matters: Coresident Sample Selection Bias in Estimating Intergenerational Mobility in Developing Countries". Email for correspondence: shahe.emran@gmail.com, fshilpi@worldbank.org.

1. Introduction

There has been a renewed interest in intergenerational economic mobility over the last few decades, with heightened concerns about widening inequality despite significant growth and poverty reduction in many developed and developing countries (World Development Report (2006), The Economist (2012)). Notwithstanding the recent interest, intergenerational economic persistence in developing countries remains an under-researched area, primarily due to data limitations.[2]

A major issue that has stunted progress in this research agenda is that the standard household surveys suffer from truncation, because coresidency is used as a criterion to define household membership (Bardhan (2014), Behrman (1999), Deaton (1997)).[3] A standard household survey such as the Living Standards Measurement Survey (LSMS) done by the World Bank, or the Household Income and Expenditure Survey (HIES) done by national statistical agencies usually includes only the coresident parents and children.[4] Since the pattern of coresidence is not random, most of the studies suffer from potentially serious sample selection bias when estimating intergenerational persistence in economic status. This has discouraged research on intergenerational economic mobility in developing countries.[5]

Although potential biases from the coresidency restriction have been a major stumbling block, to the best of our knowledge, there is no evidence on the direction and magnitude of the coresidency bias in the standard measures of intergenerational persistence in developing countries. Are the estimates from the coresident sample biased to such an extent that they

[2]The literature on intergenerational mobility in developed countries is rich with a distinguished pedigree. For excellent surveys of the literature, see Solon (1999), Black and Devereux (2011), Bjorklund and Salvanes (2011), and Corak (2013). A partial list of the contributions includes Bowles (1972), Becker and Tomes (1979), Atkinson et al. (1983), Solon (1992), Mulligan (1997), Arrow et al. (2000), Black et al. (2005), Bjorklund et al. (2006), Chetty et al. (2014), Lefgren et al. (2014), Black et al. (2015), Becker et al. (2015).

[3]Coresidency restriction results in a truncated sample, as the surveys do not gather any information on the family members who do not satisfy the coresidency criteria.

[4]Some of the children of the household head may not be part of the household at the time of the survey for a variety of reasons such as higher eduction, job, marriage and household partition.

[5]This is true even for a country such as India where there is a long tradition of high quality household survey data collection. Bardhan (2005) identifies intergenerational mobility as one of the under-researched areas of economic research in India.

are of little use in understanding intergenerational economic mobility?[6] Are the different measures of intergenerational persistence affected by coresidency bias to the same degree, or are some measures more robust than others with relatively small bias? This paper makes progress on these questions, providing evidence, analysis, and guidance for working with coresident samples that have wide ranging implications for research on intergenerational economic mobility.

To understand the implications of the coresidency restriction, the challenge is to find surveys that (i) include all of the children and the parents irrespective of their residency status, and (ii) identify the subset of individuals coresident in a household at the time of the survey. We take advantage of two high quality household surveys in villages of India and Bangladesh, and estimate the two most widely used measures of intergenerational persistence in the literature: intergenerational regression coefficient (henceforth IGRC) and intergenerational correlation (henceforth IGC).[7]

The evidence on intergenerational schooling persistence presented below in this paper shows that IGRC, the most widely used measure of intergenerational persistence, suffers from large downward bias because of truncation due to coresidency.[8] In contrast, the downward bias in the estimated IGC in coresident samples is much smaller; in many cases, less than one-third of the bias in the corresponding IGRC estimate. In the sample of 13-60 years age range, the average bias in IGRC estimates is 29.7 percent in the case of Bangladesh, while the corresponding bias in IGC estimates is only 8.7 percent. The extent of truncation bias in India is smaller because of higher coresidency rates observed in the data. However, the IGRC estimates in India are also substantially biased downward; the

[6]The prevailing view, in fact, holds that the estimates from coresident samples are not useful, and it is partly driven by the fact that the researchers do not know the direction of bias.

[7]IGRC shows how a one year of higher schooling of parents affects the schooling attainment of children. IGC shows what proportion of the variance in children's schooling can be attributed to the variance in parents' schooling. For a discussion, see Solon (1999), Hertz et al. (2007)).

[8]The bias is defined as [(Estimate from full sample - Estimate from coresident sample)/Estimate from coresident sample] $\times$ 100. A downward bias implies that the estimate from coresident sample is smaller than that from the full sample, implying a positive bias estimate according to the formula above. This is done to avoid carrying a negative sign for the bias estimates.

average bias is 17.6 percent. Again, the corresponding average bias in the IGC estimates is much smaller at 10.4 percent.[9] Considering estimates across different age ranges and gender, the average biases in IGRC and IGC are 24.4 percent and 6.5 percent respectively in Bangladesh, and the corresponding estimates in India are 14.12 percent (IGRC) and 7.6 percent (IGC). Moreover, we provide suggestive evidence that the bias in IGC estimate is less sensitive to the variations in the coresidency rate.

We discuss explanations and intuitions for the empirical findings that the IGC estimates suffer from much lower coresidency bias. A simple intuition can be provided by using the following relationship between IGRC (denoted as β) and IGC (denoted as ρ): $\rho = \beta \left(\frac{\sigma_p}{\sigma_c} \right)$, where σ_p and σ_c are standard deviations of parental schooling and children's schooling respectively. In other words, the IGC estimate is equal to the IGRC estimate multiplied by the ratio of standard deviation of parent's schooling to that of children's schooling. It is well-known that truncation biases the estimate of β downward in an OLS regression (Hausman and Wise (1977)). An equally important implication of truncation in our context is that it also affects the estimate of the ratio of standard deviations in schooling of parents to children. The IGC estimate cancels out part of the downward bias in IGRC by multiplying it with an upward biased estimate of the ratio of standard deviation of parental schooling to that of children's schooling. We also provide a plausible rationale for the finding that the IGC estimates are less sensitive to the variation in coresidency rate.

The analysis and findings presented below have important and wide ranging implications for research on intergenerational economic mobility. First, our analysis implies that much progress in understanding intergenerational mobility can be made with the household surveys available in developing countries by focusing on IGC as the measure of mobility. These data sets are currently shunned by most researchers because of the worry that the estimates from the coresident sample suffer from bias of unknown direction, and possibly of very high magnitude.[10] Second, the results in this paper can be helpful in sorting out often

[9] As we discuss later, the difference in biases between IGRC and IGC becomes smaller as the coresidency rate increases.

[10] We focus on IGRC and IGC, as they are two of the most widely used measures of mobility. But

conflicting evidence on intergenerational mobility from coresident samples; for example, in India, educational mobility has improved substantially after the 1991 reform according to the IGRC estimates, but remains largely stagnant according to the IGC estimates. Our analysis suggests that the conclusions based on the IGC estimates in such instances of conflict are more credible. Third, the evidence that the estimates are biased *downward* can be helpful in understanding changes in intergenerational mobility over time. If the estimates from coresident samples show no change or an increase in persistence over time, we can be confident that mobility has declined, as the coresidency rate in a country usually declines over time, making the downward bias in the estimates for the younger generation larger in magnitude.[11] Fourth, our results have important implications for cross-country comparisons of economic mobility. Most of the available data sets suffer from coresidency restrictions, and the extent of truncation is likely to vary across countries significantly. Since the bias in IGRC is larger, and it responds more to changes in the coresidency rate, a ranking according to IGRC is more likely to be incorrect compared to a ranking based on IGC.[12] Fifth, the evidence indicates that the IGRC estimates in coresident samples are likely to underestimate the gender gap in intergenerational economic mobility, because coresidency rates for girls are much lower in many developing countries, especially where girls leave the natal family after marriage.

The rest of the paper is organized as follows. Section 2 provides a brief discussion on the related literature, especially focusing on developing countries, and puts the contribution of this paper in perspective. The next section (section 3) discusses the data sources and variables used in the analysis. Section (4) reports the estimates of IGRC and IGC in educational attainment for Bangladesh and India data, both for the full and the coresident

our results suggest that when working with coresident samples, one should avoid measures that do not normalize for changing variances across generations.

[11] See, for example, Emran and Sun (2015) on China, Emran and Shilpi (2015) on India.

[12] The evidence below shows that, based on IGRC estimates from coresident samples, one would conclude, incorrectly, that intergenerational educational persistence is similar in India and Bangladesh, when the IGRC estimates from full samples show that persistence is substantially higher in Bangladesh. In contrast, the IGC estimates from coresident samples provide both a correct ranking, and a reliable estimate of the gap between the countries.

samples. The next section (section (5)) reports evidence from a number of alternative samples, for different age ranges of children. Section (6) provides an explanation for the findings that the coresident sample bias in the IGC estimates is small, often ignorable, especially when compared to the bias in the IGRC estimates. The following section (section (7)) discusses the implications of the results for sorting out the conflicting evidence from the existing studies on intergenerational mobility in developing countries. The paper concludes with a summary of the results and their implications for the emerging literature on intergenerational mobility in developing countries.

2. Related Literature

The literature on intergenerational economic mobility in developed countries is vast, but the corresponding literature on developing countries is limited at best. The economics literature on intergenerational mobility in developed countries has focused on intergenerational income correlations, with an emphasis on the link between fathers and sons (see, for example, Solon (1992, 1999), Mazumder (2005), Corak and Heisz (1999), Bowles et al. (2005)). The relative neglect of research on developing countries is evident from the fact that, in his survey for the handbook of labor economics, Solon (1999) cites only two papers: Lam and Schoeni (1993) on Brazil, and Lillard and Kilburn (1995) on Malaysia.

Research on intergenerational economic persistence in developing countries has been constrained primarily by two types of data limitations. First, the income data on parents and children are not available for more than a few years to allow reliable estimation of permanent income across generations. As shown by a substantial body of literature on developed countries, it is necessary to have good quality income data over a period of more than a decade to address the attenuation bias in the estimate of income persistence (Solon (1992), Mazumder (2005)). The household surveys available in developing countries usually provide income information only for a single year, and estimating individual income may be a daunting task in rural areas where self employment, work sharing, and informal activities predominate (Deaton (1997)). The second challenge which constitutes the focus

of this paper comes from the coresidency restriction; most of the surveys suffer from sample selection due to coresidency used to define household membership. As noted before, this has been a strong discouraging factor for researchers worried about rejection by their peers, journal referees and the editors.

The recent economics research on intergenerational economic mobility in developing countries includes Behrman et. al. (2001), Hertz et al. (2007), Binder and Woodruff (2002), Thomas (1996), Lillard and Willis (1995), Lam and Schoeni (1993), Emran and Shilpi (2011, 2015), Bossuroy and Cogneau (2013), Maitra and Sharma (2010)). Most of the studies on economic mobility in developing countries rely on education and occupation as markers of economic status, because reliable data on income for long enough time periods to calculate permanent income are not available.[13] Most of them also use data selected non-randomly due to the residency requirement for household membership. There is, however, no uniformity in the definitions of 'household' across different surveys, although all are concerned with 'living together', 'eating together', and sometimes with 'pooling of funds' (Deaton (1997)). Examples of household surveys that usually include coresidency as a defining criteria include Household Income and Expenditure Survey (HIES), Demographic and Health Survey (DHS), and Living Standard Measurement Survey (LSMS). There are some household surveys which include limited information on the parents of household head and spouse, but do not include the nonresident children of the household head. Hertz et al. (2007) use household surveys from 21 developing countries (10 Asian, 4 African, and 7 Latin American) and 8 formerly Communist countries where household surveys provide information on household head's parents, but do not include the nonresident children.[14] When non-resident children are excluded from the survey, it results in truncation of the

[13]The data on the income of parental generation is especially difficult to find. Preponderance of home based economic activities including own-farming in parental generation makes it challenging to estimate income in many developing countries.

[14]Hertz et al (2007) are careful about sample selection bias, and they do not focus on the household head's children as has been the case in many recent studies that rely on data without non-resident children. To the best of our knowledge, the only survey in Hertz et al. list of countries that cover all of the non-resident children in the survey is that for Bangladesh.

sample, information on both the dependent and explanatory variables for them is missing from the data set. This also implies that, in most of the cases, it is not possible to estimate a sample selection equation to correct for the biases, because it is not possible to identify if a household is missing children from the survey. The maximum likelihood approach developed by Bloom and Killingsworth (1985) can be applied in this case if multivariate normality is a reasonable assumption.

Although non-random sample selection due to coresidency has been a major methodological issue in the research on intergenerational mobility, evidence on the magnitude of coresidency bias has been scarce, with the exception of the analysis of occupational mobility in the UK by Francesconi and Nicoletti (2006). In an interesting paper, they use British Household Panel Survey to estimate the extent of coresidency bias in the estimates of intergenerational persistence in occupational prestige between father and son(s). They use the occupational prestige index due to Goldthorpe and Hope (1974), and estimate intergenerational elasticity as a measure of persistence. The evidence reported in their paper shows that the coresidency bias is substantial, ranging between 20-40 percent.[15] They, however, do not address the question whether intergenerational correlation (IGC) and intergenerational regression coefficient (IGRC) are affected differently by the truncation due to coresidency, which is the focus of our analysis.

We are not aware of any analysis of coresidency bias in the context of educational mobility, either in developed or developing countries. Our analysis can also claim broader applicability as we use data from two developing countries with substantial differences in the coresidency rates, and provide evidence on both father-son and mother-daughter links in educational persistence.

3. Data and Variables

We use two rich data sets particularly suited for the analysis of the extent of coresident

[15]They provide an extensive analysis of alternative econometric approaches for selection correction. Their findings indicate that the inverse probability weighted estimator is the most reliable to tackle coresident sample selection bias among a number of approaches including Heckman selection correction.

sample bias. The source of data on India is the 1999 Rural Economic and Demographic Survey done by the National Council for Applied Economic Research, and the data on Bangladesh comes from the 1996 Matlab Health and Socioeconomic Survey (MHSS). The Bangladesh survey collected information on household head and spouse's all children (including from past marriages) irrespective of their residency status from 4538 households in Matlab thana of Chandpur district.[16] The India survey also collected information on all of household head's children from current marriage but not non-coresident mothers of children from earlier marriage(s). We utilize these information to create data sets containing education and other personal characteristics of parents and children. Both of these surveys focus on rural areas in respective countries. An advantage of rural samples is that the bias from censoring due to possible non-completion of younger children may not be as important, because only few go on to have more than middle school (or high school) education. The children who go for more than high school education (10 years of schooling in Bangladesh and India) are also the children who leave the village household, because the "colleges" (for grades 11 and 12) and universities (for three-four year undergraduate, and graduate study) are not located in villages.

Our estimation sample consists of household head and spouse, and their children, including those from other marriages in the case of Bangladesh. For the empirical analysis, we use alternative samples defined by different age ranges for the children. Our main results are based on a sample of children aged 13-60 years. To test the sensitivity of our conclusions with respect to the specific age cutoffs, we estimate the IGRC and IGC for a number of alternative age ranges; 16-60, 20-69 and 13-50 years.

Table A.1 reports the summary statistics of the relevant variables for both the Bangladesh and India data sets for our main estimation sample (children in the age range 13-60 years). Several interesting observations and patterns are noticeable in our data sets. The average

[16]The MHSS 1996 is a collaborative effort of RAND, the Harvard School of Public Health, the University of Pennsylvania, the University of Colorado at Boulder, Brown University, Mitra and Associates and the International Centre for Diarrhoeal Disease Research, Bangladesh (ICDDR,B).

schooling attainment remains low in rural areas of both Bangladesh and India at the time of the survey years. The mean and median years of schooling are 4.97 and 5.00 respectively for Bangladesh, and 6.23 and 7 for India. The relatively lower education attainments in Bangladesh compared with India were present during parents' generation as well: median years of father's education was 2 years in Bangladesh compared with 2.50 years in India. The average number of children per household in Bangladesh is about 5.74 compared with 3.53 in India. This difference probably reflects the fact that Bangladesh data include information on children from other marriages while India data do not. There are some differences in the age distribution of children also: median age for Bangladesh data is 30 years compared with 33 years for India. The gender gap in education between boys and girls is about 1 year in Bangladesh in contrast with 2 years in India.

Table A.1 also reports the ratio of standard deviation of parent's education to that of children's education for both all and coresident children in columns 3 and 7. The ratio is unambiguously smaller in the full sample (including both coresident and non-resident children) compared with that in the coresident sample. This is consistent with the observation noted earlier in the introduction that a higher estimate of this ratio in a coresident sample is likely to partially offset the biases in IGC estimates.

Figures 1.A (Bangladesh) and 1.B (India) plot the probability of nonresidency at the time of the survey against the schooling of children. The graphs in both Bangladesh and India show that probability of nonresidence is higher in the tails. Also, the probability of nonresidence is higher for girls at any given level of schooling, although the gender gap closes substantially at the right tail in the case of Bangladesh.

4. Empirical Results

We begin the discussion with a graphical presentation of the data, following the classic analysis of truncation in Hausman and Wise (1977). Figures 2 and 3 report the bivariate linear plots of children's schooling against parents' schooling for both the full and the coresident samples for Bangladesh and India respectively. The coresidency rate is much

higher in the India data compared to that in the Bangladesh data, thus the resulting truncation bias is likely to be relatively lower in India. For example, in the father-son sample the coresidency rate is 79 percent in India, while the corresponding rate is only 52 percent in Bangladesh. In the mother-daughter samples, the coresidency rates are lower: 39 percent in India and 26 percent in Bangladesh, reflecting the fact that women leave the natal family following marriage in both countries.

For each country we present three graphs: (i) son-father, (ii) daughter-mother, and (iii) all children-father. The figures show that the slope of the fitted line is *smaller* in the coresident sample which is consistent with Hausman and Wise (1977). The widely held belief that the coresidency bias in the estimates of IGRC is substantial thus appears clearly visible in the graphs.

In the graphs for the "all children" sample (both sons and daughters), the coresident line intersects the full sample line from above (see Figures 2.A for Bangladesh and 3.A for India). This implies that the surveys miss less educated children from households with low parental education, but miss better educated children from households with high parental education. We thus have both truncation from above and from below.

A closer look at the other graphs reveals some interesting differences across gender and countries. In Bangladesh, the fitted lines in father-son sample (see figure 2.B) intersect each other at a very low level of father's education, implying that most of the coresident line lies below the full sample line. This implies that, for most of the distribution, the better educated sons leave the parental household. For Mother-daughter sample in Bangladesh (figure 2.C) the pattern of sample selection is different; the line for the coresident sample intersects the full sample line from above at about 5 years of mother's schooling which is very high given that the average education for mothers is only 1.47 years. This implies that that coresident line lies above the full sample line for most of the cases; the girls with relatively lower education leave the parental household (presumably following marriage, they relocate to husband's house). Also, the gap between the coresident and full sample lines becomes smaller as the parental education increases, which suggests that the probability of a less

educated girl leaving her parental household becomes smaller when parent's education is higher. This can be interpreted as suggestive evidence that better educated parents are less likely to marry off their daughters without completing high school (10th grade in both Bangladesh and India).

The figures for India (3.A, 3.B, and 3.C) are broadly similar, although the effect of truncation on the slope is smaller compared to the case of Bangladesh, especially in the father-son sample, which reflects the fact that the coresidency rate is very high for sons in India. However, the graphs again tell a consistent story; in all three groups, the coresident fitted line has lower slope than that in the fitted line in the full sample. The intersection points of the coresident and full sample lines are, however, more centered, implying that for the lower educated parents, it is the low educated children that leave the household, and for the high educated it is the opposite. The intersection for the daughters' is at a higher level of father's schooling, implying that the low educated daughters are non-resident for most of the cases.

While the graphical exploration provides suggestive evidence, to get a measure of the extent of bias in IGRC and IGC, we now turn to the estimates for both Bangladesh and India. We first discuss the results for the all children sample (i.e, that includes both sons and daughters). These provide average estimates across gender, and are useful as summary measures. We then provide estimates for the father-son and mother-daughter intergenerational persistence which have been the focus of most of the economics literature.

The regression specification used for estimating the IGRC and IGC is motivated by Solon (1992) and includes age and age squared of both the child and the father.[17] As robustness checks, we also estimate a number of alternative specifications, starting with a simple bivariate model where no controls are used. In addition to the quadratic age formulation standard in the literature, we use a completely flexible specification of the effects of age by including dummies for different years of age. The estimates are very robust; the numerical magnitudes of IGRC and IGC estimates vary little, if at all, across

[17]Mother's age is missing for a significant proportion of children.

different specifications.

Following the literature, we estimate the following regressions by OLS (denote IGRC by β and IGC by ρ):

$$(IGRC) \qquad S_i^c = \beta_0 + \beta S_i^p + X'\Gamma + \epsilon_i$$
$$(IGC) \qquad Z_i^c = \rho Z_i^p + \tilde{X}'\Pi + \eta_i$$

In the IGRC regression, S_i^c and S_i^p are years of schooling of children and parents respectively, and X is a set of control variables. All of the variables in IGC regression are normalized to have zero mean and unit variance; for example, we define $Z_i^c = \dfrac{S_i^c - \bar{S^c}}{\sigma_c}$, and $Z_i^p = \dfrac{S_i^p - \bar{S^p}}{\sigma_p}$, where a bar on a variable denotes sample mean, and σ_c and σ_p are the estimated standard deviations of children's schooling and parent's schooling respectively.

To help keep track of the discussion across different samples, we note here again the terminology used. We call "all children" when the sample includes both sons and daughters. A "full sample" includes both coresident and non-resident members, and "coresident sample" includes only the members coresident in the household at the time of the survey.

4.1 Estimates for All Children (Sons and Daughters)

Evidence from Bangladesh

Table 1 reports the estimates of IGRC and IGC for all children in Bangladesh data, i.e., sons and daughters combined together. The first two columns in Table 1 report the estimates of IGRC for the full and coresident samples (top panel) and the implied bias (bottom panel). We use three different measures of parental education: father's schooling, mother's schooling, and the average of father's and mother's schooling. Note that some researchers also use maximum schooling (of mothers and fathers) as a measure of parental education. In our data sets, the father has higher schooling in most of the cases, and the correlation between the maximum parental schooling and father's schooling is high enough to yield virtually identical estimates of IGRC and IGC. In addition to quadratic age

controls, we also include a dummy for gender of the child in the regression specification.[18] This implies that any common factors (such as cultural norms) that might affect the average schooling attainment of girls irrespective of parental socioeconomic status are absorbed as a shift in the intercept.

The estimates in the top panel of Table 1 provide strong evidence that truncation bias in the IGRC estimates is substantial for all three definitions of parental education. Consistent with the expectation based on the graphs discussed above, the IGRC estimate in the coresident sample is significantly biased *downward*. The null hypothesis that the estimate from the coresident sample is equal to the estimate from the full sample is rejected unambiguously with P-values equal to 0.00 in all of the different cases.[19] The pattern is remarkably consistent, and justifies the widespread opinion that there are good reasons to expect the IGRC estimates to be biased downward due to non-random sample selection bias because of the coresidency requirement used in the household surveys.

To get a better sense of the implied magnitudes, we report bias defined as follows (using IGRC as an example),

$$Bias = \frac{(IGRC_F - IGRC_{CR}) \times 100}{IGRC_{CR}}$$

where $IGRC_{CR}$ denotes the estimate from a coresident sample, while $IGRC_F$ is the estimate from the corresponding full sample including non-resident household members.

The first column in the bottom panel of Table 1 reports the bias in the IGRC estimates from the coresident sample. The evidence is clear: the estimate from coresident sample is biased downward, and the magnitude of bias is substantial across all three indicators of parental education. The bias is the highest when mother's schooling is the indicator of

[18]The estimates and the conclusions do not depend on the inclusion of the gender dummy.

[19]We, however, note here that the formal test of equality of estimates may not be very useful in our context. Even with very small numerical difference between the estimates from the full and coresident samples, one can reject the null hypothesis of equality simply because the standard errors are extremely small (see, for example, the IGC estimates). So the focus should be on the magnitude of the bias not the statistical test of equality of estimates.

parental education (34 percent), and the lowest in the case of average parental schooling (24 percent), with an average bias of 29.7 percent.[20] A 30 percent bias on an average vindicates the unease among the researchers and editors of journals that the available household surveys in developing countries may not be particularly helpful in understanding the magnitude of intergenerational persistence in economic status.

We now turn to the IGC estimates for Bangladesh reported in columns 4 (full sample) and 5 (coresident sample) of Table 1. The estimated IGCs for three different indicators of parental education are reported in the top panel and the implied biases are reported at the bottom. The evidence is strikingly different; the estimate of IGC from the coresident sample is much closer to that from the full sample, and this is true for all three different indicators of parental education (top panel). The average bias in the IGC estimates is 8.7 percent which is less than one-third of the average bias in the IGRC estimates (29.7 percent). The *highest* magnitude of bias is 11 percent in the case of IGC which is less than half of the *lowest* bias found in the IGRC estimates (24 percent).

Evidence from India

Table 2 reports estimates of IGRC and IGC from India data for three different indicators of parental education (father's schooling, mother's schooling, and average schooling of mother and father). The difference between the IGRC estimates from the coresident and full samples in the case of India are smaller in magnitude compared to the estimates from Bangladesh (compare top panel of Table 1 to that of Table 2). The average bias is about 17.6 percent. While the extent of bias is not as dramatic as in the Bangladesh data, the evidence still indicates that the coresident sample selection causes substantial downward bias in the IGRC estimates. The relatively lower selection bias in the India estimates reflects the fact that the proportion of coresident children is higher in India compared to Bangladesh (61 percent in India and about 40 percent in Bangladesh).

The IGC estimates in columns (4) and (5) in Table 2 show that the tuncation bias in

[20]It is the simple average of the three bias estimates in the bottom panel.

IGC estimates is significantly smaller. The average bias in IGC for India is 10.4 percent which is much smaller than the 17.6 percent average bias found in the IGRC estimates.

The evidence in Tables 1 and 2 thus suggest that *(i)* truncation due to the coresidency restriction in a survey causes large downward bias in the estimates of IGRC, and *(ii)* the corresponding bias in the IGC estimates is substantially lower. The widespread caution about coresidency bias seems right on target for IGRC estimates, while the IGC estimates from coresident samples are much closer to the estimates from the full samples.

India-Bangladesh Comparison

If a researcher relies on IGRC estimates from coresident samples to understand differences between Bangladesh and India in intergenerational persistence in schooling, she is more likely to reach an incorrect conclusion. For example, with father's schooling as the measure of parental education, the IGRC estimates for India and Bangladesh are very close to each other (0.42 in Bangladesh and 0.43 in India, implying a 2 percent higher estimate in India), which suggests that educational mobility is similar in the two neighboring countries. However, the results from the full sample show a different picture: a 12 percent higher estimate of IGRC in Bangladesh. In contrast, the IGC estimates from coresident samples show a 12 percent larger estimate for Bangladesh, much closer to the corresponding estimate from the full sample: a 16 percent larger estimate for Bangladesh. The IGC estimates from coresident samples thus lead to the correct ranking that educational mobility is lower in Bangladesh, and also provide a reliable measure of the gap between the two countries. The other estimates in Tables (1) and (2) also show that the IGRC estimates from coresident samples severely underestimate the gap between Bangladesh and India, while the IGC estimates yield a much more consistent and reliable picture. A broader implication of the above examples is that cross country comparisons of intergenerational mobility based on IGRC, by far the most commonly used measure, are more likely to yield incorrect conclusions, while IGC based comparisons seem much more reliable.

4.2 Estimates of Father-Son and Mother-Daughter Schooling Persistence

In this subsection, we discuss the biases in the IGRC and IGC estimates for the intergenerational link between the father and sons, and the mother and daughters. While father-son intergenerational persistence in economic status has been the most widely researched topic both in developed and developing countries, it is probably equally (if not more) important from a policy perspective to understand the barriers faced by the girls in education. The results on father-son linkage are reported in the upper panel of Table 3, and the bottom panel contains the corresponding estimates for mother-daughter persistence in schooling. We report the estimates of bias, and test the null hypothesis of zero bias (i.e., that the estimates from the coresident and the full samples are equal). For the sake of brevity, we omit the underlying estimates of IGRC and IGC. The estimates for Bangladesh are in the first two columns, and the last two columns refer to the corresponding results for India.

Bangladesh

The estimates of father-son intergenerational link in schooling for Bangladesh shows that the IGRC estimate in the coresident sample suffers from strong downward bias; the bias is 29.5 percent (row 1, column 1 in the top panel of Table 3). The bias in father-son IGRC estimate is thus similar to the average bias for the all children sample discussed above: 29.7 percent. The corresponding bias in the estimated IGC is much smaller: only 8.9 percent (row 2, column 1).

The results for mother-daughter in Bangladesh are reported in columns 1 and 2 of the lower panel of Table 3. The bias in the IGRC estimate from the coresident sample is much stronger at 45.6 percent, a very high magnitude indeed. This illustrates starkly that relying on the coresident sample can lead to a grossly misleading picture of intergenerational persistence between mother and daughter(s). This high bias reflects the fact that the degree of sample selection is very high in the daughters' case; only 26 percent of the full sample satisfies the coresidency restriction in Bangladesh data (for sons it is 52 percent of the full sample). The bias in the IGC estimate from coresident sample is again much smaller in

magnitude: 10.6 percent.

India

The estimates of father-son schooling persistence for India are reported in columns (3) and (4) of the top panel of Table 3. The IGRC estimate for India shows that the downward bias due to coresidency is substantial; the estimate from the full sample is 29.5 percent higher than the estimate in the coresident sample. The bias in the father-son sample in India is thus significantly larger than the average bias we found earlier for the sample of all children across different measures of parental education (17.6 percent). In sharp contrast, the IGC estimate suffers from very little coresidency bias: 2.4 percent only. The estimated bias in the IGC estimate for father-son in India is thus ignorable, while the IGRC estimate suffers from strong downward bias from coresident sample selection.

The bias estimates for mother-daughter schooling persistence in India are reported in columns (3) and (4) of the lower panel of Table 4. The bias in the IGRC estimate for mother-daughter is smaller for India when compared to Bangladesh, but the magnitude of bias is still substantial 21.8 percent. The corresponding biases in IGC estimates is 9.7 percent which is less than half of the bias in the IGRC estimate.

Gender Differences in Intergenerational Schooling Persistence

An important policy issue in many developing countries is whether the girls face especially strong barriers to educational mobility. If we rely on the IGRC estimates from coresident samples, the gender gap may seem smaller than it really is, because the truncation bias is usually stronger for the estimates for girls, as coresidency rates are lower (this is true in both Bangladesh and India data). But the IGC estimates from coresident samples provide a reliable measure of the gender gap. For example, consider the estimates of father-son and mother-daughter persistence in Bangladesh (Tables 3 and 5).[21] Averaging over estimates for four age ranges in Tables 3 and 5, the IGRC estimates from coresident

[21] We focus on father-son and mother-daughter links as the persistence runs along gender lines with cross effects much smaller.

samples suggest that persistence is about 33 percent higher for daughters, while the correct estimate from full samples is 45 percent. In contrast, the gender gap estimates based on IGC are similar across full (4.5 percent higher for daughters) and coresident (4.8 percent higher for daughters) samples. Thus, consistent with the cross-country comparisons discussed above, when working with coresident samples, it is preferable to use IGC as a measure of intergenerational persistence to understand gender gap in educational mobility.

5. Additional Evidence

5.1 Alternative Age Ranges for the Children

The age range used so far in Tables 1-3 is 13-60 years. This is motivated by the fact that the average schooling attainments in rural Bangladesh and India remain low in the survey years, so that a 13 years lower threshold may not be binding for most of the rural children. In Bangladesh data, the average years of schooling is only 4.43 years; for sons it is 5.5 years and for daughters 3.4 years. The average schooling in India is 5 years, and for sons it is 7 years and for daughters 3.7 years. To explore the sensitivity of the conclusions with respect to the age range of children, we estimate the IGRC and IGC across a number of different age ranges. For the sake of brevity, we report estimates from the following age ranges: (*i*) 13-50 years, (*ii*) 16-60, and (*iii*) 20-69 years.

Since many children start first grade at age 6, a 13 years age cut-off implies 7 years of potential schooling as the minimum threshold in our sample (primary schooling is 5 years). The observed schooling attainment, however, may vary across 13 year old children for a variety of reasons. For example, children from poor households may start schooling later than usual, and they may also have to interrupt schooling because of negative economic shocks.[22] The variations in schooling attainment at age 13 (or even younger) can thus provide us useful evidence on the role played by family background. However, one might worry that some children at age 13 have not yet completed schooling, and it is important

[22]According to one estimate for India, 53 percent of students drop out before completing primary (5 years). Among every 100 girls enrolled, only 40 progress to 4th grade, 18 reaches 8th grade, and only 1 is lucky enough to go up to 12th grade (India Education Report, 2005, pp. 6-7).

to check if the results hold when the lower threshold for children's schooling is raised. We thus estimate the IGRC and IGC in a series of samples, starting with 14 years and raising the lower threshold incrementally by one year at a time up to 20 years. The evidence from this exercise is very reassuring; while the magnitudes differ across the samples, the main conclusions reached on the basis of 13-60 years age range remain intact. For the sake of brevity we report estimates for 16 and 20 years as the lower age threshold for children. A 16 year age cut-off implies potentially 10 years of schooling which coincides with one of the most important public examination in both Bangladesh and India (called Secondary School Certificate (S.S.C) or 'Matriculation' examination). After 12 years of schooling (18 years of age cut-off), the students sit for a second important public examination, called Higher Secondary Certificate (H.S.C) or Intermediate examination. In our Bangladesh data, about 10 percent of 20 years of age or older has 10 years or more schooling, and 5 percent has 12 years or more schooling.

For each age range, we present the bias estimates and omit the underlying estimates of IGRC and IGC for full and coresident samples. This allows us to reduce the number of tables by putting together the relevant estimates for both Bangladesh and India in a single table. However, all of the underlying estimates are available from the authors upon request. Note that we do not discuss statistical significance of the estimates, all of the estimates are significant at the 5 percent or lower level. But as noted earlier, the estimated standard errors in IGRC and IGC are very small reflecting the large sample size. Thus statistical significance is not very informative, and the focus should be on the magnitude of the bias.

The estimates for three different age ranges for the all children sample including both sons and daughters are presented in Table 4. The estimates, both for India and Bangladesh, lead to the same set of conclusions derived from the 13-60 years age range in section (4). The IGRC estimates, in general, suffer from substantial downward bias because of truncation due to coresidency. In Bangladesh, the bias in the coresident sample estimate of IGRC is more than 10 percent in seven out of nine cases, with an average bias of 15.56 percent. The extent of bias in the case of India is smaller, the average is 14.7 percent, but still the bias

is more than 10 percent in seven out of nine cases.

In contrast, the coresidency bias in the IGC estimates are again much smaller, in only one out of nine cases the bias is more than 10 percent in Bangladesh (10.9 percent when father's age is the indicator of parental education and the age range for children is 13-50 years). The average bias in the IGC estimate for Bangladesh is only 4.6 percent, less than one-third of that in the IGRC estimates (15.56 percent). The estimates for India are similar; in three out of nine cases the bias is more than 10 percent, the highest being 12 percent, and the average is 9 percent, significantly smaller than the corresponding average for IGRC (14.7 percent).[23]

Table 5 reports estimates for the father-son and mother-daughter persistence in schooling attainment for different age ranges of children. As to be expected, the magnitudes of the estimates vary across different age ranges, but the main conclusions of the paper remain valid. The coresidency bias in the IGRC estimates is very high in the estimates for Bangladesh; the lowest bias is 15 percent and the highest 46 percent, with an average of 27 percent, a very high bias by any standard. The bias estimates for India are smaller in magnitude consistent with its higher coresidency rates. However, the average bias in IGRC is still more than 10 percent (10.6 percent). *More important for the research on intergenerational mobility in developing countries constrained by the coresident samples is the clear evidence that in all 12 cases reported in Table 5, the bias in the IGC estimate is much smaller than that in the corresponding IGRC estimate. The average bias in the IGC estimates is only 7.5 percent in Bangladesh (27 percent for IGRC), and 4.5 percent in India (10.6 percent for IGRC).*

5.2 Coresidency Rates and the Extent of Bias

An interesting aspect of the results presented above is that there is significant variation

[23]Note that there are two negative estimates of bias out of a total of 36 estimates in Table 4, implying that the estimate from full sample is smaller than that from the coresident sample. It is, however, important to underscore that, in both of these cases, the numerical magnitudes of estimates from full and coresident samples are extremely close; the IGRC estimates are 0.83 (full sample) and 0.84 (coresident sample), and the IGC estimates are 0.53 (full sample) and 0.54 (coresident sample).

in the coresidency rates across Bangladesh and India data, and the bias estimates reflect the differences in the severity of selection. Since we estimated the biases in IGRC and IGC for a number of different samples, one might wonder how the magnitude of the bias relate to coresidency rate across different samples. Figure 4 shows the relation between the coresidency rate and the estimated bias for both IGRC and IGC estimates. There is a clear negative relation between coresidency rate and the magnitude of bias in the case of IGRC, implying that comparing IGRC estimates from different data sets may not be appropriate. In contrast, there is no discernible relation between the bias in IGC estimates and the coresidency rate. An OLS regression of the bias in IGRC estimates on coresidency rates yield a coefficient of -0.22 which is significant at the 1 percent level (t statistic equals -2.50). The coefficient on a regression of the bias in IGC estimates on the coresidency rates is, in contrast, numerically very small (0.008) and statistically insignificant ($t = 0.30$ and P-value= 0.77). We provide an explanation of the low sensitivity of IGC estimates to coresidency rates below (please see section 6.2).

6. Toward an Understanding of the Results: Why Is the Bias in IGC Estimates So Low?

The evidence presented above is strikingly consistent and clear: when a researcher works with a data set from a survey that uses coresidency for defining the household, the IGRC estimates are likely to be seriously biased downward; but the estimates of IGC in coresident samples are, in general, much closer to the estimates from the full sample.

It is important to appreciate that the coresident sample bias common in the household surveys in developing countries is best modeled as a truncation, not censoring. The most common problem in the context of household surveys in developing countries is that there is no information (on both dependent and independent variables) for the the non-resident children resulting in truncation of the sample. The evidence presented above suggests that the non-resident children are not randomly distributed, both in Bangladesh and India: they come mostly from the tails of the schooling distribution.

6.1 Coresidency Restriction and Truncation Bias in a Simple Model

6.1.1 Bias in the IGRC Estimate

Consider the standard model of sample truncation widely discussed in the econometrics and statistics literature, adapted to our application (for the econometric literature see Heckman (1976), Greene (2012), and for a statistical treatment, see Cohen (1991)). The truncation is from below and based on a level of schooling of the children $T > 0$; so a girl i with schooling level $S_i^c \leqq T_i$ leaves the household for marriage, for example, and thus is not included in the survey. A simple model of the marriage decision is as follows (assuming parent's decide marriage for girls):

$$M_i = \begin{cases} 1 \quad if \qquad v_i - wS_i^c > 0 \\ 0 \qquad\qquad otherwise \end{cases} \tag{1}$$

where v_i is payoff (indirect utility) from marrying off child i , wS_i^c is the labor market earnings forgone as a girl leaves the natal family after marriage, and S_i^c is the schooling level of girl i. The marriage decision M_i is a binary indicator that takes on the value of 1 when a girl is married (and lives in a separate household).

Denote the set of individuals included in the survey by D. So child i is unmarried and thus coresident with the parents and is included in the survey, i.e., $i \in D$, if the following holds:

$$S_i^c > \frac{v_i}{w} \equiv T_i$$

So we have the following model of the population relation and data generation:

$$S_i^c = \beta_0 + \beta S_i^p + \epsilon_i;\ i \in D,\ if\ S_i^c > T_i > 0 \tag{2}$$

where S_i^p denote years of schooling of parents. We assume that $\epsilon_i \sim N\left(0, \sigma_c^2\right)$.

For simplicity of exposition, we ignore other control variables X such as age of parents and child. A standard result in the literature is that OLS regression in the coresident

sample suffers from omitted variables bias, because the conditional expectation function is not linear (Greene (2012), Heckman (1976)):

$$S_i^c = \beta_0 + \beta S_i^p + \frac{\sigma_{v\epsilon}}{\sigma_v}\lambda_i + \epsilon_i \tag{3}$$

where $\sigma_{v\epsilon}$ is the covariance between v_i and ϵ_i, and σ_v is the standard deviation of v_i.

The error term in the OLS regression is not ϵ_i, but $\mu_i = \frac{\sigma_{v\epsilon}}{\sigma_v}\lambda_i + \epsilon_i$ which is correlated with S_i^p causing omitted variables bias. The omitted variable λ_i is called the inverse Mills ratio and given as follows:

$$\lambda_i = \lambda(\alpha_i) \equiv \frac{\phi(\alpha_i)}{1 - \Phi(\alpha_i)} \quad \alpha_i = \frac{T_i - \beta_0 - \beta S_i^p}{\sigma_c}$$

As discussed by Greene (2012), although the bias depends on the correlations in the data, a robust empirical regularity widely observed in the literature is that the OLS estimate is biased downward to zero (see also Hausman and Wise (1977), Cohen (1991)). Hausman and Wise (1977) discuss a rationale for the downward bias by showing that the OLS estimate is necessarily smaller than the maximum likelihood estimate (see the appendix to Hausman and Wise (1977)).

Denoting the OLS estimate in the coresident sample by $\hat{\beta_T}$, the attenuation bias due to truncation in the OLS estimate can be approximated by the following relationship:[24]

$$plim\left(\hat{\beta_T} - \beta\right) \approx (\delta - 1)\beta < 0 \tag{4}$$

where

$$\delta = \left[1 - \alpha\lambda(\alpha) - (\lambda(\alpha))^2\right] \in (0, 1)$$

and α is the mean of α_i. Our estimates of IGRC $\left(\hat{\beta_T}\right)$ for Bangladesh and India show that the bias implied by inequality (4) above can be serious.

[24]See Greene (2012) for a more complete discussion on this.

6.1.2 Bias in the IGC Estimate

The IGC can be estimated from a regression where the variables are normalized so that their mean is zero and variance is 1. Denote the IGC (correlation coefficient) between father's schooling and children's schooling by ρ. The we have the following regression model for estimation of IGC:

$$Z_i^c = \rho Z_i^p + \sigma_{\eta\vartheta}\widetilde{\lambda}_i + \eta_i \quad i \in D,\ if\ Z_i^c > \tilde{T}_i \equiv \left(\frac{T_i - \bar{S}^c}{\sigma_c}\right) \tag{5}$$

where

$$\begin{aligned} Z_i^c &= \frac{S_i^c - \bar{S}^c}{\sigma_c} & Z_i^p &= \frac{S_i^p - \bar{S}^p}{\sigma_p} \\ \widetilde{\lambda}_i &= \lambda\left(\tilde{\alpha}_i\right) & \tilde{\alpha}_i &= \tilde{T} - \rho Z_i^p \\ \eta_i &= \frac{\epsilon_i}{\sigma_c} & \vartheta_i &= \frac{v_i}{\sigma_c} \end{aligned}$$

As noted earlier, a bar on a variable denotes the sample mean, and σ_c and σ_p are the standard deviations of children's and parental schooling respectively, and $\sigma_{\eta\vartheta}$ is the covariance between the error terms in the children's schooling and marriage selection equation with the schooling variables standardized. The truncation point in the standardized model is $\tilde{T}_i = \dfrac{T_i - \bar{S}^c}{\sigma_c}$.

To see that the truncation bias is lower in OLS estimate of equation (5), note that similar to equation (4) above, we have the following approximate relation for model (5):

$$plim\left(\hat{\rho}_T - \rho\right) \approx \left(\widetilde{\delta} - 1\right)\rho < 0 \tag{6}$$

where

$$\widetilde{\delta} = \left[1 - \tilde{\alpha}\lambda\left(\tilde{\alpha}\right) - \left(\lambda\left(\tilde{\alpha}\right)\right)^2\right]$$

It is easy to check that $\widetilde{\delta} > \delta$, if $\tilde{\alpha} < \alpha$. By using the relation that $\beta = \rho\dfrac{\sigma_p}{\sigma_c}$, we can

rewrite $\tilde{\alpha}_i$ as follows:

$$\tilde{\alpha_i} = \left(\frac{T - \beta_0 - \beta S_i^p}{\sigma_c}\right) - \left(\frac{\bar{S^c} - \beta_0 - \beta \bar{S^p}}{\sigma_c}\right) = \alpha_i - \left(\frac{\bar{S^c} - \beta_0 - \beta \bar{S^p}}{\sigma_c}\right) \tag{7}$$

Now $\tilde{\alpha}_i < \alpha_i$ follows from the observation that $\left(\bar{S^c} - \beta_0 - \beta \bar{S^p}\right) > 0$ in a truncated sample because $\bar{S^c} = E\left(S_i^c \mid S_i^c > T_i\right) = \beta_0 + \beta \bar{S^p} + E\left(\epsilon_i \mid \epsilon_i > T_i - \beta_0 - \beta S_i^p\right)$
and $E\left(\epsilon_i \mid \epsilon_i > T_i - \beta_0 - \beta S_i^p\right) > 0$.

6.2 Discussion

The preceding section provides a conceptual basis for the empirical evidence from Bangladesh and India presented in the earlier part of this paper. Here we discuss alternative ways to think about the coresidency bias in the IGC and IGRC estimates which may provide additional intuitions.

We focus on the following relationship between IGRC and IGC widely known in the literature (see, for example, Solon (1999)):

$$\rho = \beta \frac{\sigma_p}{\sigma_c} \tag{8}$$

A simple way to understand the evidence presented in this paper is that truncation biases the estimate of β downward, but it also results in upward bias in the estimate of ratio of standard deviations in schooling $\frac{\sigma_p}{\sigma_c}$. As a result, the net bias in IGC (ρ) is smaller than the bias in IGRC (β) estimate. Estimate of the ratio of the standard deviations in our data sets confirms that the magnitude is larger in the truncated samples (see Table A.1).

A standard result from the literature is that truncation reduces the variance of a variable (Greene (2012)). Since truncation is based on children's schooling, it affects the variance of children's schooling directly :

$$Plim\left(\hat{\sigma}_c\right) = \sqrt{\delta}\left(\sigma_c\right) \tag{9}$$

Note that the commonly available household surveys in developing countries include a random sample of parents (household head and spouse), and thus the estimate of the standard deviation of parental schooling is likely to be unbiased. We can put together the relations in inequality (4), and equations (8) and (9) to derive the following approximate relation:

$$Plim\left(\hat{\rho}_T - \rho\right) \approx \left(\sqrt{\delta} - 1\right)\rho \tag{10}$$

Now observe that $\sqrt{\delta} > \delta$, because $\delta \in (0, 1)$, and as a result, the bias represented by the right hand side of approximation (10) is much smaller than the bias in approximation (4). To give a sense of the magnitudes, $\delta = 0.9$ implies a value of $\sqrt{\delta} = 0.949$, and $\delta = 0.8$ implies $\sqrt{\delta} = 0.90$. Thus the IGC estimates from coresident sample suffer from much less bias when compared to the most widely used measure of intergenerational persistence: IGRC. If the bias in IGRC is 10 percent, the corresponding bias in IGC is half of that (5 percent), and when the IGRC estimate is biased downward by 20 percent, the corresponding bias in IGC is about 10 percent. An important implication of the above results is that, at high levels of coresidency rates (δ closer to 1), the difference between the biases in IGRC and IGC will be smaller, which is consistent with the evidence that in India the differences in biases between IGRC and IGC is much smaller when compared to that in Bangladesh. The actual biases estimated in the data, however, also reflect sampling variability.[25]

From relations (4) and (10) above, we get the following approximate results on the slope of the bias in IGRC and IGC estimate with respect to the coresidency rate:

$$\begin{aligned} \frac{\partial Plim\left(\hat{\beta}_T - \beta\right)}{\partial \delta} &\approx 1 \\ \frac{\partial Plim\left(\hat{\rho}_T - \rho\right)}{\partial \delta} &\approx \frac{1}{2\sqrt{\delta}} \end{aligned} \tag{11}$$

The results in equation (11) above are important for cross-country and over time (for

[25] Our evidence that in many cases, the bias in IGC is less than a third of the corresponding bias in IGRC implies that these approximate relations underestimate the advantages of IGC as a measure of mobility. However, we believe that this is useful in providing an intuitive understanding of the empirical results.

a given country) comparisons of intergenerational mobility, because they suggest that the bias in IGC estimates responds less with variations in coresidency rate when compared to the IGRC estimates, assuming that coresidency rates are not too low ($\delta > 0.25$).[26] This provides a conceptual basis for the striking differences in the slopes of IGRC and IGC estimates with respect to coresidency rates in figure 4 above. Thus when coresidency rates vary across countries or over time, the IGC estimates are likely to provide us with a more accurate ranking of countries and evolution of economic persistence over time.

7. Implications for the Existing Studies and the Debate on Economic Mobility in Developing Countries

In the introduction, we briefly mentioned that it is not uncommon in the literature to find that conclusions regarding intergenerational mobility in economic status in developing countries depend on the measure used. A survey of the literature shows that the studies that rely on IGRC as the metric, in general, conclude that economic mobility has increased substantially over time (see, for example, Jalan and Murgai (2008) on India, and Hertz et al. (2007) for cross-country evidence). The evidence based on IGC on the other hand tend to find much more stickiness in social mobility, and conclude that mobility has not improved in any significant way (Emran and Shilpi (2015) on India, and Hertz et al. (2007) for cross-country evidence).

Hertz et al. (2007), using a sample of 42 countries (21 of them developing countries), report a sustained and significant decline in the magnitudes of the estimated IGRC in schooling over time. They also report IGC estimates which show a very different picture: there is no discernible trend in the estimates; the slope of the fitted line is, in fact, close to zero. Hertz et al. (2007) are very much aware of the critical role played by the differences in the variance in schooling across generations; they emphasize the fact that the variance of children's schooling relative to the variance of parent's schooling has gone down over the years in the data they use, and that decline explains the divergence between the IGRC and

[26]The lowest coresidency rate across different samples in our analysis is 26 percent, for the mother-daughter sample in Bangladesh.

IGC estimates. They, however, do not note the possible connection between the variance of the children's schooling and the truncation bias due to coresidency restriction in the survey, as the educated children are more likely to move out of parental home in the younger generations, because of improved labor market opportunities, increased geographic mobility of labor, and changes in cultural norms about age at marriage, and extended family (in favor of nuclear family) in many developing countries. It is highly likely that at least part of the declining variance may reflect sample truncation due to coresidency criterion used in surveys. A related point relevant for cross country comparisons is that the coresidency bias in the IGRC estimates is likely to vary across different countries which would depend on a variety of economic and cultural factors such as labor market opportunities for children, costs of housing, availability of public welfare schemes for ageing poor parents, among other things. As we discussed in section (6.2) above, the magnitude of bias in the IGRC estimate vary substantially with the coresidency rate, but the bias in the IGC estimate is much less sensitive (see also figure 4). An immediate and important implication of this observation is that one should be cautious about the IGRC estimates for cross-country comparison of economic mobility, the focus instead should be on the estimates of IGC.

Similar evidence can be found in recent studies on intergenerational mobility in other developing countries. Consider the case of India as an example. The extent of and trend in economic mobility in India has attracted attention of the researchers given the evidence that economic liberalization might have contributed to increased inequality while it has led to growth in income and poverty reduction. The existing estimates of intergenerational educational persistence in India lead to opposing conclusions depending on whether IGRC or IGC is used as a measure; persistence has gone down substantially according to the IGRC estimates, but it has remained largely unchanged in recent decades according to the IGC estimates (Maitra and Sharma (2010), Jalan and Murgai (2008), Emran and Shilpi (2015)). These studies focus on the parents (household head and spouse) and his/her children, and the data used in all of these studies use the coresidency restriction in the survey definition of household membership. The evidence presented in this paper implies that the

widely discussed improvements in educational mobility in India in last few decades should be interpreted with due caution because they are based on IGRC estimates from coresident samples, and thus are likely to be substantially biased downward, and the apparent improvements in IGRC may largely (or at least partly) be driven by declining coresidency rates in younger generations.

8. Conclusions

We take advantage of two rich data sets from Bangladesh and India to explore the direction and magnitude of coresident sample selection bias in the two most widely used measures of intergenerational persistence: intergenerational regression coefficient (IGRC) and intergenerational correlation (IGC). The evidence reported in this paper shows that the worry about coresidency bias is well-justified when the focus is on estimating IGRC, by far the most popular measure among development economists.[27] The IGRC estimates, in general, suffer from substantial downward bias in coresident samples vindicating the skepticism among researchers and journal editors about the usefulness of data with coresidency restriction. The bias in IGC estimates is, however, much smaller in magnitude, less than one-third of that in the IGRC estimates on an average. We discuss theoretical explanations and intuitions behind the empirical results. The biases in both IGC and IGRC converge to zero as coresidency rate converges to 100 percent, but, even with high coresidency rates, IGC estimates are preferable.

Robust evidence on the direction of bias, i.e., that the estimates of intergenerational persistence are highly likely to be biased downward in coresident samples can be useful in understanding changes in economic mobility over time, given that coresidency rates are usually lower in the younger generations. This adds a caveat to the optimistic picture of intergenerational educational mobility found in some countries such as India based on declining IGRC estimates from coresident samples over time.

Our analysis also suggests that the IGC estimates are much less sensitive to the vari-

[27] The same negative conclusion is likely to hold for other related measures of mobility where the focus is on a the slope parameter of a regression without normalization to take into account changes in variances.

ation in coresidency rates compared to the IGRC estimates. Since coresidency rates can vary substantially across countries, over time, and across gender, the IGC estimates are likely to be more reliable for understanding the pattern and evolution of intergenerational mobility. The evidence shows that the IGRC estimates from coresident samples lead to the incorrect conclusion that intergenerational schooling persistence is virtually same in India and Bangladesh. In contrast, the IGC estimates from coresident samples yield the correct conclusion that persistence is higher in Bangladesh, and also provide a reliable estimate of the gap between the two countries. The evidence from both Bangladesh and India shows that coresidency rates are lower for girls, and thus the persistence estimates suffer from stronger downward bias, which may generate a false impression of lower gender gap.

The evidence and analysis in this paper thus provide a strong rationale for focusing on IGC as a measure of intergenerational mobility in the context of developing countries. Perhaps, the most important implication of our analysis is that a large number of good quality household surveys in developing countries that use coresidency to define household membership (for example, LSMS and HIES) are not worthless in analyzing the strength, pattern and evolution of intergenerational economic persistence. Much progress could be made with the imperfect data if the researchers move away from the current emphasis on IGRC and use IGC as the appropriate measure instead.

References

Arrow, K, S. Bowles, S. Durlauf (2000). *Meritocracy and Economic Inequality*, Princeton University Press.

Atkinson, A.B., A.K. Maynard, and C.G. Trinder (1983). *Parents and Children: Incomes in Two Generations.* London: Heinemann Educational Books.

Bardhan, P (2014), The State of Indian Economic Statistics: Data Quantity and Quality Issues, Mimeo, Berkeley, CA.

Bardhan, P (2005), "Theory and Empirics in Development Economics", Economic and Political Weekly, August, 2005.

Becker, Gary S. and Nigel Tomes (1979), "An equilibrium theory of the distribution of income and intergenerational mobility", Journal of Political Economy 87:1153-1189.

Becker, Gary, & Kominers, Scott Duke & Murphy, Kevin M. & Spenkuch, Jörg L., 2015. "A Theory of Intergenerational Mobility," MPRA Paper 66334, University Library of Munich, Germany.

Binder, Melissa and Christopher Woodruff. 2002. "Inequality and Intergenerational Mobility in Schooling: The Case of Mexico." Economic Development and Cultural Change, Vol. 50, Iss. 2, pp. 249-267.

Behrman, J., A. Gaviria and M. Szekely (2001), "Intergenerational Mobility in Latin America," Economia, Vol. 2 (1): 1 44.

Behrman, Jere R., (1999). "Labor markets in developing countries", in: O. Ashenfelter & D. Card (ed.), Handbook of Labor Economics, edition 1, volume 3, chapter 43, pages 2859-2939 Elsevier.

Bjorklund A and K. Salvanes. (2011). "Education and Family Background: Mechanisms and Policies, Handbook in the Economics of Education vol 3, E A Hanushek, S Machin and L Woessmann (es.), The Netherlands: North Holland, 2011, pp. 201-247.

Björklund, A., Lindahl, M., & Plug, E. (2006). "The origins of intergenerational associations: Lessons from Swedish adoption data." The Quarterly Journal of Economics, 999-1028.

Bossuroy, T and Denis Cogneau, 2013. "Social Mobility in Five African Countries," Review of Income and Wealth, vol. 59, pages S84-S110, October.

Black, S, & Paul J. Devereux & Petter Lundborg & Kaveh Majlesi, 2015. "Poor Little Rich Kids? The Determinants of the Intergenerational Transmission of Wealth," NBER Working Papers 21409, National Bureau of Economic Research, Inc.

Black, S. E. and P. Devereux (2011). "Recent Developments in Intergenerational Mobility, Handbook of Labor Economics, Amsterdam, North-Holland.

Black, Sandra E., Paul J. Devereux and Kjell G. Salvanes (2005), "Why the apple does not fall far: Understanding intergenerational transmission of human capital," American

Economic Review 95: 437-449.

Bloom, David E., and Killingsworth, Mark R. (1985), "Correcting for Truncation Bias caused by a Latent Truncation Variable", Journal of Econometrics, 1985, pp. 131-135.

Bowles, Samuel (1972), "Schooling and inequality from generation to generation", Journal of Political Economy 80: 219-251.

Chetty, R, and N. Hendren, P. Kline, and E. Saez, (2014). "Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States," The Quarterly Journal of Economics, Oxford University Press, vol. 129(4), pages 1553-1623.

Cohen, A (1991), *Truncated and Censored Samples: Theory and Applications*, CRC Press.

Corak, M and A. Heisz (1999), "The Intergenerational Earnings and Income Mobility of Canadian Men: Evidence from Longitudinal Income Tax Data, Journal of Human Resources. Volume 34, Number 3 (Summer), pages 504-533.

Deaton, A (1997), *The analysis of household surveys: A microeconometric approach to development policy.* Oxford University Press.

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

Emran, M. Shahe and F. Shilpi (2015). "Gender, Geography and Generations : Intergenerational Educational Mobility in Post-reform India", World Development, Vol. 72, 362-380.

Emran, M. Shahe and Yan Sun (2015), Magical Transition? Intergenerational Educational and Occupational Mobility in Rural China: 1988-2002, World Bank Policy Research Working Paper 7316.

Greene, W (2012), Limited Dependent Variables - Truncation, Censoring, and Sample Selection, Chapter 19, *Econometric Analysis*, Pearson.

Hausman, J and D. Wise (1977), "Social Experimentation, Truncated Distribution, and Efficient Estimation", Econometrica, May 1977.

Heckman, J (1976), " The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models", Chapter in Sanford Berg Ed. *Annals of Economic and Social Measurement*, Volume 5, number 4.

Hertz Tom, T. Jayasundera, P. Piraino, S. Selcuk, N. Smith and A. Veraschagina (2007). "The Inheritance of Educational Inequality: International Comparisons and Fifty-Year Trends. The B.E. Journal of Economic Analysis and Policy (Advances), 7(2), Article 10.

Jalan, J. and R. Murgai, (2008). "Intergenerational Mobility in Education in India, Manuscript, World Bank, Delhi.

Lam, David, and Robert F. Schoeni. 1993. "Effects of Family Background on Earnings and Returns to Schooling: Evidence from Brazil." Journal of Political Economy 101(4):710 40.

Lefgren, L, & Matthew J. Lindquist, & David Sims, 2012. "Rich Dad, Smart Dad: Decomposing the Intergenerational Transmission of Income," Journal of Political Economy, University of Chicago Press, vol. 120(2), pages 268 - 303.

Lillard, Lee and Robert Willis. 1995. "Intergenerational Educational Mobility, Effects of Family and state in Malaysia". The Journal of Human resources, Vol. (29), pp 1126-1166.

Maitra, P and A. Sharma (2010), "Parents and Children: Education Across Generations in India, Working paper, Monash University.

Mazumder, Bhashkar (2005), "Fortunate Sons: New Estimates of Intergenerational Mobility in U.S. Using Social Security Earnings Data," Review of Economics and Statistics, May, 2005.

Mulligan, Casey B. (1997). *Parental Priorities and Economic Inequality.* Chicago: University of Chicago Press.

Solon, Gary, 1992. "Intergenerational Income Mobility in the United States," American Economic Review, American Economic Association, vol. 82(3), pages 393-408, June.

Solon, Gary (1999). "Intergenerational Mobility in the Labor Market, in O. Ashenfel-

ter and D. Card (eds.), Handbook of Labor Economics 3A, Elsevier, Amsterdam, North Holland.

The Economist (2012), "For Richer, For Poorer", Special Report on Inequality by Zanny Minton Beddoes, October 13th 2012.

Thomas, D (1996). "Education across Generations in South Africa", American Economic Review, American Economic Association, vol. 86(2), pages 330-34, May.

World Development Report (2006), Equity and Development, Oxford University Press.

Table 1: Intergenerational Persistence and Coresident Sample Bias: Bangladesh

(All Children)

	Intergenerational Regression Coefficients (IGRC)			Intergenerational Correlations(IGC)		
	Full	**Co-resident**	Test of Equality (χ^2)	**Full**	**Co-resident**	Test of Equality (χ^2)
Father's Education	**0.55*** **	**0.42*** **		**0.51*** **	**0.46*** **	
(Standard Error)	(0.0149)	(0.0134)	91.30	(0.0138)	(0.0146)	9.16
Observations	14,017	5,599		14,017	5,599	
Mother's Education	**0.86*** **	**0.64*** **		**0.47*** **	**0.44*** **	
(Standard Error)	(0.0245)	(0.0200)	140.23	(0.0134)	(0.0138)	11.01
Observations	14,527	5,523		14,527	5,523	
Parent's Education (average)	**0.73*** **	**0.59*** **		**0.52*** **	**0.48*** **	
(Standard Error)	(0.0172)	(0.0167)	132.83	(0.0123)	(0.0136)	22.55
Observations	18,505	5,806		18,505	5,806	

Biases	**Bias (IGRC)**	Co-residency Rate	**Bias (IGC)**
Father's Education	**31%**	40%	**11%**
Mother's Education	**34%**	38%	**7%**
Parent's Education (average)	**24%**	31%	**8%**

NOTES:

(1) * significant at 10%; ** significant at 5%; *** significant at 1%

(2) The Bias is defined as [(Estimate from Full Sample – Estimate from Coresident Sample) *100] / Estimate from Coresident Sample

Table 2: Intergenerational Persistence and Coresident Sample Bias: India (All Children)

	Intergenerational Regression Coefficients (IGRC)			Intergenerational Correlations(IGC)		
	Full	**Co-resident**	Test of Equality (χ^2)	**Full**	**Co-resident**	Test of Equality (χ^2)
Father's Education	**0.49*** **	**0.43*** **		**0.44*** **	**0.41*** **	
(Standard Error)	(0.0120)	(0.0134)	109.23	(0.0108)	(0.0128)	38.87
Observations	14,877	9,132		14,877	9,132	
Mother's Education	**0.57*** **	**0.47*** **		**0.37*** **	**0.33*** **	
(Standard Error)	(0.0178)	(0.0172)	88.47	(0.0115)	(0.0121)	27.62
Observations	14,877	9,132		14,877	9,132	
Parent's Education (average)	**0.66*** **	**0.56*** **		**0.46*** **	**0.42*** **	
(Standard Error)	(0.0152)	(0.0164)	54.56	(0.0106)	(0.0123)	23.41
Observations	14,877	9,132		14,877	9,132	

Biases	**Bias (IGRC)**	Co-residency Rate	**Bias (IGC)**
Father's Education	**14%**	61%	**7%**
Mother's Education	**21%**	61%	**12%**
Parent's Education (average)	**18%**	61%	**10%**

NOTES: (1) * significant at 10%; ** significant at 5%; *** significant at 1%

(2) The Bias is defined as [(Estimate from Full Sample – Estimate from Coresident Sample) *100] / Estimate from Coresident Sample

Table 3: Intergenerational Persistence between Father- Sons and Mother-Daughters (Bangladesh and India)

	Bangladesh		**India**	
	Bias	Test of No Bias (χ^2)	**Bias**	Test of No Bias (χ^2)
Father-Son Persistence				
Intergenerational Regression Coefficient	**30%**	69.54	**9%**	49.59
Intergenerational Correlation	**9%**	7.95	**2%**	9.18
Coresidency Rate	52%		79%	
Mother-Daughter Persistence				
Intergenerational Regression Coefficient	**46%**	69.54	**24%**	31.75
Intergenerational Correlation	**11%**	7.95	**13%**	8.27
Co-residency Rate	26 %		39%	

Table 4: Robustness Checks for Different Age Ranges: All Children

	Bangladesh		India	
	Intergenerational Regression Coeff (IGRC)	Intergenerational Correlations (IGC)	Intergenerational Regression Coeff (IGRC)	Intergenerational Correlations (IGC)
Biases	**Bias**			
16-60 Year Age group				
Father's Education	**20%**	**6%**	**11%**	**7%**
Mother's Education	**24%**	**4%**	**17%**	**11%**
Parent's Education (average)	**11%**	**4%**	**16%**	**9%**
20-69 Year Age group				
Father's Education	**9%**	**4%**	**8%**	**4%**
Mother's Education	**11%**	**2%**	**18%**	**11%**
Parent's Education (average)	**-2%**	**-2%**	**12%**	**7%**
13-50 Year Age group				
Father's Education	**31%**	**11%**	**14%**	**10%**
Mother's Education	**34%**	**7%**	**21%**	**12%**
Parent's Education (average)	**24%**	**6%**	**16%**	**10%**

NOTES: (1) Bias is defined as [(Estimate from Full Sample – Estimate from Coresident Sample) *100] / Estimate from Coresident Sample

(2) All of the Bias Estimates are Significant at the 5 Percent or Lower Level.

Table 5: Robustness Checks for Different Age Ranges: Father-Son, and Mother-Daughter

	Biases			
	Bangladesh		India	
	Father-Son	Mother-Daughter	Father-Son	Mother-Daughter
16-60 Year Age group				
Intergenerational Regression Coeff. (IGRC)	**22%**	**32%**	**6%**	**16%**
Intergenerational Correlations(IGC)	**9%**	**4%**	**2%**	**7%**
20-69 Year Age group				
Intergenerational Regression Coeff. (IGRC)	**15%**	**18%**	**6%**	**7%**
Intergenerational Correlations(IGC)	**6%**	**6%**	**2%**	**4%**
13-50 Year Age group				
Intergenerational Regression Coeff. (IGRC)	**30%**	**46%**	**7%**	**22%**
Intergenerational Correlations(IGC)	**9%**	**11%**	**2%**	**10%**

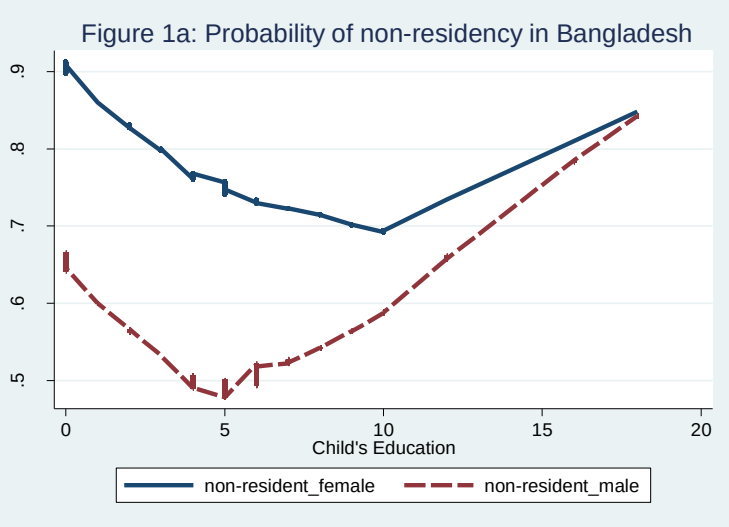
NOTE: (1) Bias is defined as [(Estimate from Full Sample – Estimate from Coresident Sample) *100] / Estimate from Coresident Sample

(2) All of the Bias Estimates are Significant at the 5 Percent or Lower Level

Table A.1: Summary Statistics

	All Children				Co-resident Children			
	Mean	**Median**	**σ_p/σ_c**	**N**	**Mean**	**Median**	**σ_p/σ_c**	**N**
				BANGLADESH				
Years of Education of				**Both Sons and Daughters Sample**				
Children	4.97	5.00		18587	5.52	5.00		5852
Father	3.39	2.00	0.92	14017	3.74	3.00	1.09	5599
Mother	1.46	0.00	0.55	14527	1.81	0.00	0.69	5523
Average of Parents	2.33	1.00	0.70	18505	2.78	2.00	0.82	5806
				Sons Sample				
Children	5.84	5.00		9056	5.56	5.00		3873
Father	3.38	2.00	0.85	7126	3.53	2.00	1.01	3713
Mother	1.45	0.00	0.51	7261	1.64	0.00	0.62	3648
Average of Parents	2.34	1.00	0.65	9010	2.59	1.50	0.75	3844
				Daughters Sample				
Children	4.14	4.00		9531	5.44	5.00		1979
Father	3.41	2.00	1.05	6891	4.16	3.00	1.27	1886
Mother	1.47	0.00	0.63	7266	2.14	0.00	0.83	1875
Average of Parents	2.33	0.50	0.80	9495	3.14	2.50	0.96	1962
				INDIA				
Years of Education of				**Both Sons and Daughters Sample**				
Children	6.23	7.00		14877	6.97	8.00		9132
Father	4.37	2.50	0.91	14877	4.74	5.00	0.95	9132
Mother	1.83	0.00	0.65	14877	2.12	0.00	0.70	9132
Average of Parents	3.10	2.50	0.70	14877	3.43	2.50	0.74	9132
				Sons Sample				
Children	7.29	8.00		8341	7.54	8.00		6561
Father	4.31	2.50	0.92	8341	4.59	5.00	0.96	6561
Mother	1.82	0.00	0.66	8341	1.99	0.00	0.70	6561
Average of Parents	3.06	2.50	0.71	8341	3.29	2.50	0.74	6561
				Daughters Sample				
Children	4.87	5.00		6536	5.54	6.00		2571
Father	4.46	2.50	0.96	6536	5.14	5.00	0.97	2571
Mother	1.84	0.00	0.68	6536	2.45	0.00	0.75	2571
Average of Parents	3.15	2.50	0.73	6536	3.79	3.25	0.78	2571

Figure 1: Child's Education and his/her probability of non-residency in Bangladesh and India



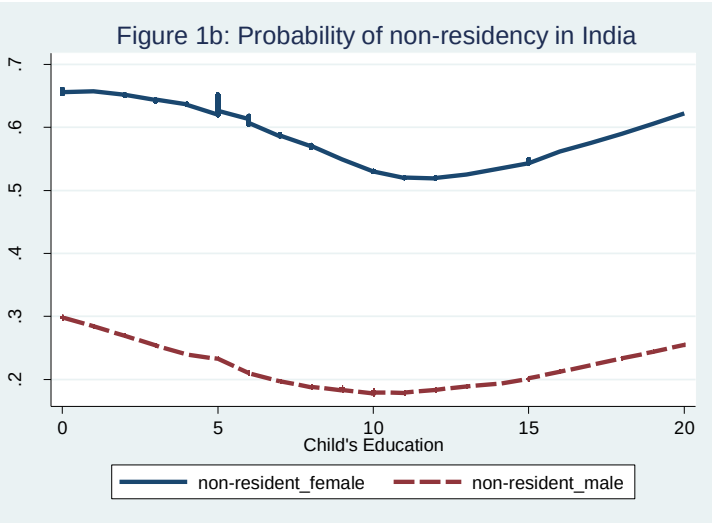
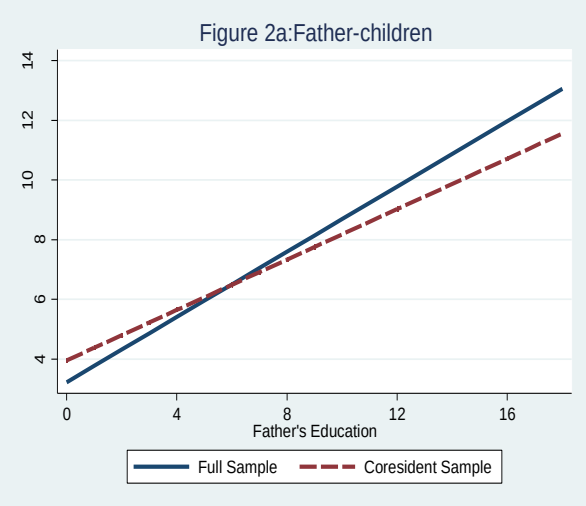
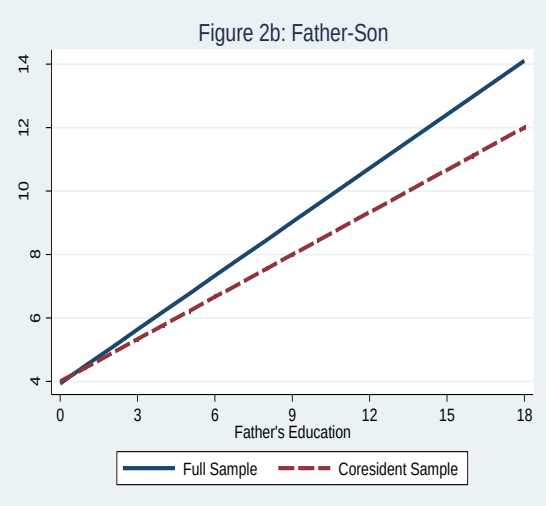


Figure 2: Fitted lines between parent's and children's education in Bangladesh (Coresident and Full Samples)





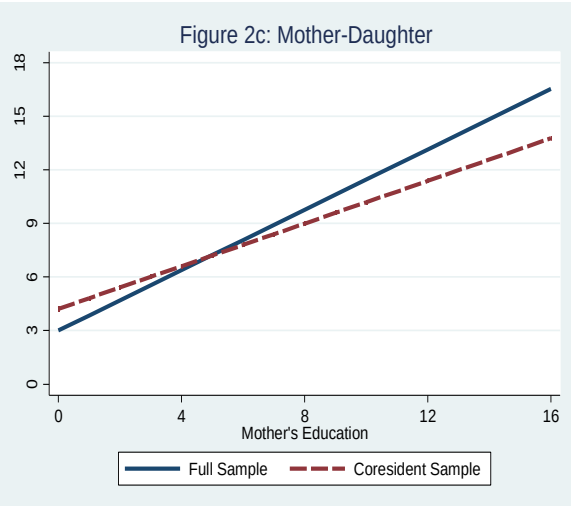


Figure 3: Fitted lines between parent's and children's education in India (Coresident and Full Samples)

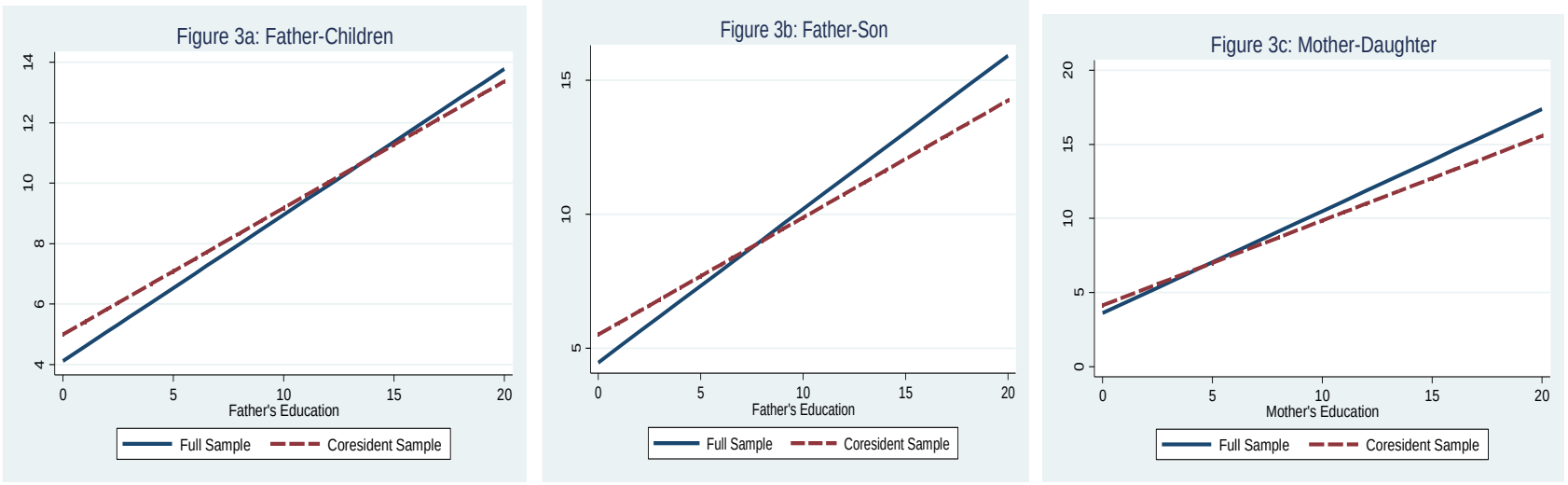


Figure 4: Co-residency Rate and Biases in Estimates of IGRC and IGC