When Measure Matters: Coresidence Bias and Intergenerational Mobility Revisited*

Ercio Muñoz†
CUNY and the World Bank

Mariel Siravegna‡
Georgetown University

This draft: December 2021

Abstract

Measurement of intergenerational mobility (IGM) in education requires linked information about children's and parents' educational attainment. However, several economies do not offer better data alternatives to estimate IGM than the use of coresident samples (i.e., samples with this link only available for individuals living with their parents), which may yield biased estimates. In this line, a recently published paper concludes that the intergenerational correlation coefficient is less biased than the intergenerational regression coefficient as a measure of relative IGM in the context of developing countries, and researchers should move away from using the latter. We re-examine this conclusion and offer empirical evidence against it. In addition, we use two data sources for 18 countries to provide evidence of the extent of coresidence bias on an extensive set of IGM indicators of absolute mobility, relative mobility, and movement. We compare estimates with retrospective information using a social survey against those obtained with coresident samples using census data for the same countries and birth cohorts. We show that there are indicators with varying coresidence bias going from less than 1% to more than 10%. Still, some mobility indicators with minimal bias produce high levels of re-ranking that make them uninformative to rank economies across time and space by the level of IGM. In contrast, other indicators with large bias generate more reliable rankings.

JEL-Codes: D63, I24, J62.
Keywords: *Intergenerational mobility, Education, Coresidence bias, Truncation bias, Coresidency, Survey data, Census data.*

---

*We thank James Albrecht, Shahe Emran, John Rust, and Susan Vroman for helpful comments on a previous draft. Marcelo Gomez provided research assistance. The usual disclaimers apply. First draft: November 2021.

†Email: emunozsaavedra@gc.cuny.edu.
‡Email: mcs92@georgetown.edu.

I Introduction

Intergenerational mobility (IGM) in education studies the relationship between the educational attainment of children and their parents. It aims to provide insights into the transmission of socioeconomic advantages in society and the degree of equality of opportunity in the economy. In particular, if the society shows a strong association between children and parents' academic outcomes, it could mean that the family's educational resources determine the success or failure of children in school. On the contrary, if the society shows a weak association, it could mean that everyone has similar opportunities to succeed regardless of their family background. From a policy point of view, it is interesting to compare country estimates to shed some light on the potential determinants or policies that influence IGM.

Several economies do not offer better data alternatives than the use of coresident samples to estimate IGM (i.e. samples with the link between parents and children are only available for those individuals who are coresiding).[1] Moreover, some data sources such as population censuses provide advantages in terms of geographical disaggregation and historical coverage but only allow the use of individuals living with parents at the time of interview (i.e. coresidence samples). Researchers are cautious about the suitability of coresident samples to measure IGM because of a potential sample selection issue. Although intuitively the problem is clear, the literature documenting the size and consequences of the bias is relatively scarce (see for example, Emran, Greene, & Shilpi, 2018; Emran & Shilpi, 2018; Francesconi & Nicoletti, 2006).

In this paper, we contribute to the understanding of intergenerational mobility in education by studying the impact of coresidence bias on its measurement. Our first contribution is to show that the correlation coefficient is not always less biased by coresidence than the regression coefficient as recently concluded (see Emran et al., 2018). We use the same simple model of coresidence analyzed by the authors and highlight the key assumption needed

[1] For example, Narayan et al. (2018); Van der Weide, Lakner, Gerszon Mahler, Narayan, and Ramasubbaiah (2021) generates estimates of IGM for 153 countries, where 39 of them are coresident samples.

for such a conclusion. Then, we discuss how pooling a large set of birth cohorts to study coresidence bias favors the correlation measure in the evidence presented to supports that conclusion. Finally, we offer new empirical evidence against the conclusion based on the two previous points.

Our second contribution is to provide novel empirical evidence of the extent to which coresident samples produce biased estimates for a large set of IGM indicators used in the literature. We compare estimates of these indicators for the same countries and same birth cohorts using two sources of data: 1) Latinobarometro social survey, that contains retrospective information about the education attainment of parents (i.e., each individual is asked the highest education attained by her parents), and 2) coresident samples obtained from census data where we link individuals aged 21-25 years to their parents only if they live together. We find average biases that go from less than 1% to more than 10%. In both absolute and relative mobility, we find indicators with small bias (close to 1%); however, some of the indicators of relative mobility with small bias also show a very small rank correlation (i.e., dissimilar ranking between sources). We also document that for some indicators this is the case even in absence of coresidence bias. Our findings suggest that the information content they provide to rank different populations across time and space according to relative mobility is very noisy. In contrast, some of the indicators of absolute mobility provide rank correlations between sources as high as 0.91, which suggests that they are very informative to rank populations even in the presence of coresidence bias. Our results in the second part of this paper have at least three implications for the recent literature on intergenerational mobility in education. First, in the case of relative mobility, the information content available in the Pearson correlation coefficient and rank-based indicators computed with education data seem to be less reliable to rank economies than the intergenerational regression coefficient despite their smaller coresidence bias. Second, researchers still need to be careful about comparisons across economies that pool indicators computed with coresident samples and those that use all children. Nonetheless, some indicators are more likely to allow such comparisons as they show very small level of coresidence bias while others are less likely because of large bias. Third, the use of coresident samples obtained from census data to study absolute mobility (as done in several recent papers) using the likelihood of achieving at least primary education conditional on parents not achieving that level provide reliable information (small bias and meaningful rankings).

The rest of the paper is structured as follows: Section II provides a brief overview of related literature, putting our contribution in context. Section III re-examines the conclusion that coresidence bias impacts the Pearson correlation coefficient less than the regression coefficient. Section IV provides empirical evidence of the extent of coresidence bias in a larger set of indicators. Finally, in Section V we conclude with some final remarks.

II Related Literature

An extensive body of literature estimates intergenerational socioeconomic mobility using different measures of status (e.g., income, occupation, education, among others) at the country-level or within countries. The research that documents IGM in income is mainly focused on high-income economies. In contrast, IGM in education's papers are predominant on developing countries (see Emran & Shilpi, 2019; Torche, 2019, for recent surveys). In general, this divergence is in part driven by differences in the type of data available in these countries.

In terms of measurement, there is variety of indicators being used. Deutscher and Mazumder (2021) recently provides a framework to classify these different measures of intergenerational mobility in income into five main groups: 1) global measures of relative mobility; 2) local measures of mobility; 3) global measures of absolute mobility; 4) global measures of movement, and 5) broad measures of relative mobility. A similar mapping can be applied to the indicators used in the literature of IGM in education.[2] Table 1 describes a (non-exhaustive) set of indicators that can be found in recent articles on IGM in education

[2]Discussions about the type of indicators in the literature of IGM in education can also be found in Narayan et al. (2018); Neidhöfer, Serrano, and Gasparini (2018); Torche (2019).

grouped into three categories: 1) Absolute mobility: including global measures as the share of children with higher education than parents (see YOS, CAT, and MIX) and local measures based on conditional probabilities that focus on particular segments of the population (see BUM-primary, BUM-secondary, TDM-primary, TDM-secondary, and UCP in Table 1); 2) Relative mobility: including global measures such as the intergenerational regression coefficient, intergenerational correlation coefficient and rank correlation (see IGRC, IGPC, and IGSC in Table 1), and local measures such as the conditional expected rank or rank-based transition probabilities (see CER050 and BHQ4); and 3) Movement: that considers global indicators of movement based on Fields and Ok (1996) and a variant used in Van der Weide et al. (2021) that can be considered a local measure of movement (see M1, M2, and DIF in Table 1).

Table 1: Indicators of Educational Intergenerational Mobility

Name	Description
Absolute Mobility	
YOS	Share of children with more years of schooling than parents, $YOS = Pr(S^y > S^o \mid S^o < max(S^o))$
CAT	Share of children with higher level of education than parents, $CAT = Pr(C^y > C^o \mid C^o < max(C^o))$
MIX	A variant of CAT such that $MIX = Pr(C^y > C^o \quad or \quad C^y = C^o = max(C^o))$
BUM-primary	Bottom upward mobility: $Pr(C^y \geq primary \mid C^o < primary)$
BUM-secondary	Bottom upward mobility: $Pr(C^y \geq secondary \mid C^o < secondary)$
TDM-primary	Top down mobility: $Pr(C^y \geq primary \mid C^o < primary)$
TDM-secondary	Top down mobility: $Pr(C^y \geq secondary \mid C^o < secondary)$
UCP	Upper class persistence: $Pr(C^y \geq secondary \mid C^o \geq secondary)$
Relative mobility	
IGRC	OLS estimate of the slope (β) in $S^y = \alpha + \beta S^o$
IGPC	Pearson correlation coefficient (ρ), where $\rho = Corr(S^y, S^o)$
IGSC	Spearman correlation coefficient, $IGSC = Corr(R^y, R^o)$
CER050	Expected rank of children with parents in bottom half, $CER050 = \mathbb{E}(R^y \mid R^o \leq 50)$
BHQ4	Prob. of reaching top quartile if parents are in bottom half, $BHQ4 = Pr(R^y > 75 \mid R^o \leq 50)$
Movement	
M1	Average change in schooling between generations, $M1 = \frac{1}{N}\sum \lvert S_i^y - S_i^o \rvert$
M2	Average directional change in schooling between generations, $M2 = \frac{1}{N}\sum (S_i^y - S_i^o)$
DIF	Same as M2 but for children with parents that did not complete tertiary

Notes: S^y and S^o denotes years of schooling of children and parents, respectively. C^y and C^o denotes educational attainment as categories (e.g., 1=less than primary, 2=primary, 3=secondary, and 4=tertiary) for children and parents, respectively. R^y and R^o denotes percentile ranks computed using years of schooling of children and parents, respectively.

In terms of data, a non-negligible share of the estimates in the recent literature rely on coresident samples as the information to link children's educational attainment to one of their parents is not always available. For example, Table 2 provides a summary of data and indicators in several recent studies that use coresident samples.

There are three things to highlight from this set of papers. First, there is a novel interest in exploring intergenerational mobility within countries. For example, Alesina, Hohmann, Michalopoulos, and Papaioannou (2020, 2021); Asher, Novosad, and Rafkin (2021); Card, Domnisoru, and Taylor (2018); Dodin, Findeisen, Henkel, Sachs, and Schüle (2021); Munoz (2021a); Van der Weide, Ferreira de Souza, and Barbosa (2020) focus on a sub-national level. Second, several studies seek to build indicators that allow comparisons across countries and/or regions (see for example, Alesina et al., 2020; Munoz, 2021a). An important implication of this is that different samples need to be comparable, and the ranking that results from pooling the indicators from all these sources needs to be meaningful. Third, all these studies focus on a small number of birth cohorts that are observed at young ages at the time of the interview. This is done to minimize potential coresidence bias by focusing on individuals at an age that is old enough to complete a given level of education but young enough that the majority still coreside with their parents. Moreover, most of the authors using census data rely on measures such as bottom upward mobility (e.g., the likelihood of completing at least primary education conditional on having parents who did not complete that level), focusing on a level that can be completed at a young age. The use of census data is related to the interest in sub-national measures and the fact that household survey data typically do not allow this type of analysis because of sample size and limitations in representativeness.

As we mentioned before, the literature addressing the consequences of using coresident samples in the context of intergenerational mobility is relatively scarce. To the best of our knowledge, there are three papers focused directly on the issue (Emran et al., 2018; Emran & Shilpi, 2018; Francesconi & Nicoletti, 2006), which we summarize in what follows. Francesconi and Nicoletti (2006) look at occupational intergenerational mobility in the UK with data from the British Household Panel Survey and find evidence that the magnitude of the bias is substantial. Emran et al. (2018) analyze coresidence bias in the context of two indicators of relative intergenerational mobility concluding that the intergenerational

Table 2: Recent literature using coresident samples to estimate IGM in education

Article	Coverage	Data and Sample	Indicators
Alesina et al. (2021)	Africa	69 censuses (aged 14-25)	BUM, TDM
Alesina et al. (2020)	Africa	37 censuses and 1 hh. survey (aged 14-18)	BUM, TDM
Asher et al. (2021)	India	2011-12 SECC Census (aged 20-23)	BUM, TDM (interval)
Card et al. (2018)	US	Census 1940 (aged 14-18 and 14-16)	BUM
Derenoncourt (2021)	US	Census 1940 (aged 14-18)	BUM
Dodin et al. (2021)	Germany	Microcensuses 1997-2018 (aged 17-21)	IGIG, Q5/Q1, Q1
Feigenbaum (2018)	Iowa	Census 1915 Iowa and 1940 US (aged 3-17)	IGRC
Geng (2020)	China	Census 1982, 1990, and 2000 (aged 23-32)	IGRC, IGPC, IGSC
Hilger (2016)	US	Censuses from 1940 to 2000 (aged 26-29)	IGRC, IGRI
Munoz (2021a)	LAC	96 censuses (aged 14-25)	BUM, TDM
Munoz (2021b)	Chile	Census 2017 (aged 21-25)	IGRC, YOS
Van der Weide et al. (2021)	153 countries	Household surveys (aged 21-25)	YOS, CAT, IGRC, IGPC
Van der Weide et al. (2020)	Brazil	Census 2010 (aged 20-24)	IGRC, IGPC, YOS, IGRI

Notes: A description of most of the indicators (BUM, TDM, IGRC, IGPC, IGSC, YOS, and CAT) can be found in Table 1. IGRI corresponds to the intercept in a regression between years of schooling of children against those of parents. Dodin et al. (2021) use some variations of the measures discussed here that combine information of educational attainment with income (income gradient, BUM ratios). LAC refers to Latin America and the Caribbean region. Van der Weide et al. (2021) also uses MIX, DIF, CER050, and BHQ4 for robustness and only 39 out of their 153 samples use coresidents.

correlation is less biased than the intergenerational regression and suggest that researchers should move away from the latter. The authors provide evidence from survey data in India and Bangladesh to support this conclusion. Finally, Emran and Shilpi (2018) assess how coresidence bias affects rank-based mobility estimates relative to intergenerational regression coefficient and intergenerational correlation. The authors conclude that the bias in rank-based absolute mobility estimates is the lowest in most cases, which suggests that this measure is the most suitable for this type of research.

We are not aware of any previous analysis of coresidence bias in the context of educational mobility looking at the following two factors: 1) to what extent coresidence bias affects a large set of indicators as used in the recent literature, particularly the bottom upward mobility often used with census data, and 2) to what extent the coresidence restriction produces re-ranking of the populations under analysis. This last point is different from the size of the bias, given that researchers could use a group of biased estimates to rank economies across time and space if the bias is large but does not vary significantly across these populations.

III IGRC versus IGPC

We start our analysis of coresidence bias by reassessing the main conclusion put forward by Emran et al. (2018), i.e., that the intergenerational correlation coefficient suffers less from coresidence bias than the intergenerational regression coefficient. With this purpose, we re-state these conclusions using the same simple model of coresidence. Then, we reassess the validity of these conclusions in the specific context in which coresident samples have been recently used (see Table 2) and discuss how the empirical evidence that supports their conclusion is constructed favors the correlation over the regression coefficient. Finally, we use household survey data with retrospective information to provide novel evidence supporting our main points.

III.1 Coresidence bias in the simple model of Emran et al. (2018)

To motivate the missing data scheme in the context of IGM, consider a set of individuals D included in a survey. In this model, parents (denoted by o) take the marriage decision for their own children (denoted by y). For instance, if a child gets married, she will leave the house; otherwise, she will stay home. Suppose the children get married and they do not live at home with their parents. In that case, the information about their level of education will not be available in the survey, truncating the sample. The marriage decision (M_i) is modeled as a binary indicator that takes values of 1 if the child gets married and 0 otherwise:

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

According to the equation 1, a child with the level of education S_i^y will get married if the indirect utility (v_i) of her progenitors from marrying off their child is greater than the labor market earnings generated if the child stays at home (wS_i^y). Otherwise, if the child is unmarried, her information is included in the survey, and the following equation holds:

$$S_i^y > \frac{v_i}{w} \equiv T_i \tag{2}$$

Hence, the underlying econometric model for the estimation of the intergenerational regression coefficient (IGRC=β) is the following linear regression equation:

$$S_i^y = \beta_0 + \beta S_i^o + \epsilon_i \quad i \in D, \quad \epsilon_i \sim N(0, \sigma_y^2), \quad if \quad S_i^s > T_i > 0 \tag{3}$$

Given the coresidence restriction, the error term has two parts:

$$S_i^y = \beta_0 + \beta S_i^o + \underbrace{\beta_v \lambda_i + \mu_i}_{\epsilon_i} \tag{4}$$

where λ_i corresponds to the inverse Mills ratio and $\beta_v = \frac{covariance_{v,\epsilon}}{variance_v}$ (i.e., relationship between the payoff from marrying off a child and her level of schooling) and the the structural error ϵ_i. If this is the case, $\mathbb{E}(\epsilon_i|S_i^o) \neq 0$, which will mean that there is omitted variable bias. As Emran et al. (2018) explain in their paper, this formula gives us a simple way to determine the sign of bias. If the indirect utility of marrying off a child and the child's level of education are positively correlated, the bias is downward (i.e., $plim(\hat{\beta} - \beta) < 0$). Nonetheless, the authors assume downward bias based on a empirical regularity observed in the literature.

In the case of the intergenerational Pearson correlation coefficient (IGPC=ρ), it can be written as:

$$\rho = \beta \frac{\sigma_{S^o}}{\sigma_{S^y}} \tag{5}$$

where σ_{S^o} and σ_{S^y} are the standard deviations in years of schooling for the sample of parents and children, respectively.

Emran et al. (2018) concludes that the intergenerational correlation coefficient is less biased and hence more robust to coresidence bias than the intergenerational regression coefficient. The intuition is simple; as the simple model shows, OLS has a downward bias for the estimation of β, but the ratio $\frac{\sigma_{S^o}}{\sigma_{S^y}}$ has an upward bias. Hence, these two biases in opposite directions play in favor of ρ. The idea that the ratio of standard deviations has an upward bias comes from the fact that S^y is truncated (which implies lower variance) and the assumption that S^o is likely unbiased because the household survey sample includes a random sample of household heads and spouses.

The authors offer empirical evidence to support the conclusion that ρ is less biased than β using two household surveys with data from India and Bangladesh, where household heads are asked about the level of education of all their children regardless of their coresidency status. This evidence is based on a sample of children aged 13-60 years but includes some sensitivity analysis with age ranges: 16-60, 20-69, and 13-50 years.

III.2 Is IGPC less biased than IGRC? A re-examination

We make two simple points regarding the previous analysis that make the conclusion that IGPC is less biased than IGRC unwarranted. First, the assumption that the ratio of standard deviations $\frac{\sigma_{S^o}}{\sigma_{S^y}}$ has upward bias is unlikely to hold in the setup in which recent papers are done. Moreover, the IGRC may not be necessarily biased downward either, which suggests that the relative impact of coresidence bias is an empirical question more than a theoretical one.[3] Second, the empirical evidence presented in Emran et al. (2018) pooling approximately five decades of children's birth cohorts, which may favor the correlation coefficient given the documented fact that the correlation coefficient tends to be more stable across cohorts. In what follows, we discuss these two points in detail.

The bias of $\sigma_{S^o}/\sigma_{S^y}$ is not necessarily upward. The reasoning behind the assumption that the ratio of standard deviations has upward bias relies on the idea that household surveys randomly select household heads and spouses and ask their educational attainment. Therefore, we can use them to estimate the variance in the schooling of parents without

[3]We focus on the ratio but in section IV we also present empirical evidence that the bias is upward on average.

bias. This is certainly true; however, researchers typically estimate the correlation coefficient using the set of complete cases (i.e., observations where children and parents education are available). Therefore, they also estimate the standard deviation of schooling using a truncated sample of parents (e.g., a head that is parent but do not have any children currently living at home is not used in the estimation of the standard deviation), which is likely to be truncated in the same direction as the sample of children given the positive correlation between parents and children educational attainment. This implies that the bias depends on the relative magnitude of the truncation in both samples (parents and children) and that in some cases, IGPC may be even more biased than IGRC if the ratio has a bias in the same direction as IGRC.

As shown in Table 2, several recent papers use census data to estimate intergenerational mobility. When this is the case, researchers typically restrict the sample to children born in a small number of years such that they are old enough to complete their education but young enough to coreside with their parents. We could also argue that in such setup, the standard deviation of schooling for children can be estimated without bias as we could observe all of them. However, as we argue in the previous paragraph, researchers typically use complete cases to estimate the correlation coefficient and therefore, use truncated samples.

Given the previous discussion, we believe that the sign of the bias in the case of the ratio of standard deviations cannot be assumed to be in one particular direction ex-ante, and it may vary across places or cohorts. We will show how the bias indeed varies across cohorts for one specific country. Moreover, in the next section, we will offer additional empirical evidence that indeed varies across different samples using information from 18 countries.

Pooling a large number of birth cohorts may favor IGPC in bias comparisons. Emran et al. (2018) use data from India and Bangladesh to show that the bias in the case of the IGRC is larger than with the IGPC. The main evidence is a comparison of estimates of both indicators using the information of all children aged 13-60 years and then only the sub-sample that coresides with their parents.

Our second point is that the comparison of bias is done by pooling a large number of birth cohorts favors the indicator with lower variation across cohorts, which happens to be the IGPC. In what follows, we explain why this is the case.

Without loss of generality, consider that there are 2 cohorts with different levels of intergenerational mobility such that:

$$S^y_{ic} = \alpha_c + \beta_c S^o_{ic} + \epsilon_{ic} \qquad i \in [1, N_c] \qquad c = 1, 2 \tag{6}$$

where we assume ϵ_{ic} is independent of S^o_{ic} and c denote cohorts. However we estimate the model pooling these cohorts.

In this framework, to assess the magnitude of the coresidence bias using pooled cohorts we would estimate an OLS regression pooling all the information and using all the children to get the following estimate as the benchmark:

$$\hat{\beta}^{pooled} = \frac{\sum_{i=1}^{N_1}(S^y_{i1} - \bar{S}^y)(S^o_{i1} - \bar{S}^o) + \sum_{i=1}^{N_2}(S^y_{i2} - \bar{S}^y)(S^o_{i1} - \bar{S}^o)}{\sum_{i=1}^{N_1}(S^o_{i1} - \bar{S}^o)^2 + \sum_{i=1}^{N_2}(S^o_{i2} - \bar{S}^o)^2} \tag{7}$$

where $\bar{S}^y = \frac{\sum_{i=1}^{N_1} S^y_{i1} + \sum_{i=1}^{N_2} S^y_{i2}}{N_1+N_2}$ and $\bar{S}^o = \frac{\sum_{i=1}^{N_1} S^o_{i1} + \sum_{i=1}^{N_2} S^o_{i2}}{N_1+N_2}$.

This benchmark estimate, under the assumption that ϵ_{ic} in uncorrelated to parents schooling within and across cohorts, has the following expected value:

$$\begin{aligned} \mathbb{E}[\beta^{pooled}] = \beta_1 & \frac{\sum_{i=1}^{N_1}(S^o_{i1} - \bar{S}^o)^2}{\sum_{i=1}^{N_1}(S^o_{i1} - \bar{S}^o)^2 + \sum_{i=1}^{N_2}(S^o_{i2} - \bar{S}^o)^2} \\ +\beta_2 & \frac{\sum_{i=1}^{N_2}(S^o_{i2} - \bar{S}^o)^2}{\sum_{i=1}^{N_1}(S^o_{i1} - \bar{S}^o)^2 + \sum_{i=1}^{N_2}(S^o_{i2} - \bar{S}^o)^2} \\ & = \beta_1 W_1 + \beta_2 W_2 \end{aligned} \tag{8}$$

Equation 8 means that β^{pooled} can be interpreted as a weighted average of the level of IGRC faced by our 2 cohorts. These weights are somewhat arbitrary given that they consider the share of variation in schooling of parents (pooling all cohorts) accounted by each cohort.

An equivalent derivation (omitted for the sake of brevity) can be constructed for the IGPC given that it can be computed using a regression like in equation 6 with standardized years of schooling, which give us that:

$$\mathbb{E}[\rho^{pooled}] = \rho_1 \tilde{W}_1 + \rho_2 \tilde{W}_2 \tag{9}$$

where $\tilde{W}_1$ and $\tilde{W}_2$ are similar weights based on the squared deviation from the mean using standardized years of schooling of parents.

Given that coresidence rates vary with age (younger people coreside with parents at higher rates), even if coresidence conditional on age is fully random, the weights for each cohort in a coresident sample will vary (relative to the benchmark that uses all children), assigning less weight to older cohorts (because they have lower coresidence rates). Hence, even if we were able to estimate intergenerational mobility with coresident samples without bias for each cohort (or age group) separately, the pooled estimate using all the cohorts with the coresident sample will likely be biased due to the change in weights. Moreover, something to note is that this is not a problem if the indicator of intergenerational mobility does not change across cohorts (i.e., $\beta_1 = \beta_2$ in our example). Hence, this will likely favor the IGPC given the documented fact that, in general, it varies less than the IGRC across cohorts (see for example, Hertz et al., 2007). Figure A1 in the Appendix shows that this is true in the case of India, where IGRC shows a pronounced decline since 1940 while IGPC has remained relatively flat.[4]

III.3 Empirical evidence

Data. We use the year 2013 wave of a nationally representative household survey called Encuesta Nacional de Calidad de Vida (ENCV) from Colombia. The survey collects information about the educational attainment of all the members of each household interviewed

[4]Unfortunately, the data source does not have estimates across cohorts for Bangladesh.

and additionally ask about the educational attainment of the father and mother of these members and whether they are coresiding with father and mother.

Results. Table 3 reports the main empirical evidence supporting the two points made in the previous section. We estimate two indicators of intergenerational mobility (IGRC and IGPC) for different age groups and pool all these age groups together. We do so using all children and only the coresident sample. We also estimate the ratio of the standard deviation in years of schooling of children over the one of parents respectively. In addition, we report the size of the coresidence bias (difference between estimates with full sample versus coresident sample), sample sizes and the coresidence rate of each age group.

Several findings emerge from Table 3. First, we find that in the case of Colombia, the level of intergenerational mobility has been declining when measured with the IGRC but stays relatively stable when measured with the IGPC (see the top two rows). This is also observed when the coresident sample is used, and it matches the general pattern discussed in the previous section that is also observed for India. Second, we find that the IGRC and IGPC are downward biased in all the age groups and the pooled group with the exception of the oldest age group (56-65). However, when the magnitude of the bias is compared, a striking pattern emerges.

When we compared different age groups, the bias typically favors IGRC (see ages 56-65, 46-55, 36-45, and 21-25), but when we pool all age groups, it favors IGPC. Even more strikingly, the bias in the IGRC computed pooling all the cohorts is more than double the size of the highest bias found for one particular age group. Third, the ratio of standard deviations is not always upwardly biased as assumed in Emran et al. (2018). In our data set, age groups 56-65, 26-35, and 21-65 are biased upward, while the other three age groups are biased downward.

In the Appendix (see Tables A1 and A2), we show that very similar patterns emerge when we replicate Table 3 using household survey data from Ecuador and Guatemala, although in the case of Guatemala the ratio of variances is indeed biased upward for all the age groups.

This rules out that these results may be related to some specificity of Colombia.

Taking all of the previous findings together, the empirical evidence supports the idea that pooling different age groups or birth cohorts may severely increase the bias in the estimates of the IGRC for reasons other than the coresidence restriction itself (i.e., other than the potential correlation between children's education and their coresident status) and that the ratio of standard deviation may not always show upward bias. Hence, we conclude that researchers should not discard estimates of the IGRC in favor of the IGPC as previously suggested.

Table 3: Coresidence bias for two relative indicators of intergenerational mobility in Colombia's ENCV 2013 household survey

	Age groups (children)					
	21-25	26-35	36-45	46-55	56-65	21-65
IGRC	.39	.47	.55	.63	.69	.56
IGPC	.52	.53	.51	.53	.53	.56
IGRC (coresident sample)	.39	.44	.54	.61	.71	.49
IGPC (coresident sample)	.51	.5	.49	.48	.55	.54
Bias in IGRC (%)	-.32	-7.2	-1	-4	3.3	-13
Bias in IGPC (%)	-.56	-5.6	-3.9	-9	3.7	-3
Ratio of SD (σ_p/σ_c)	1.3	1.1	.94	.83	.77	1
Ratio of SD (coresident sample)	1.3	1.1	.91	.79	.77	.91
Bias in ratio of SD (%)	-.24	1.7	-2.9	-5.3	.41	12
N	5368	9599	8598	7654	5048	36267
Coresidence rate (%)	53	31	17	11	6.5	23

Notes: The table reports estimates of the intergenerational regression coefficient (IGRC) and the intergenerational Pearson correlation coefficient (IGPC) computed for different age groups and pooling all these groups. These estimates use years of schooling (YOS) censored at 15 for children and their parents. The latter uses the highest level when information about both parents are available. We use all the children of a given age and then restrict the sample to those that coreside with at least one parent (i.e., coresident sample). We report the bias in these indicators as a percentage of the value computed with the full sample (coresidents and no coresidents). We report the standard deviation (SD) of YOS of parents (σ_{S^o}) over the SD in YOS of children (σ_{S^y}), and the bias computed as a percentage of the ratio estimated with the entire sample. The row N reports the number of children used in the estimation with the full sample, and the coresidence rate in the last row indicates the percentage of all children (i.e., N) living with at least one parent.

IV Coresidence bias in a larger set of indicators

In this section we expand our focus to include the full set of absolute mobility, relative mobility, and movement indicators described in Table 1. In particular, we compare the estimates of each indicator for the same country and birth cohorts computed with a data source containing retrospective information (individuals are asked about their parents' education) against those obtained with a data source that only contains information for individuals living with their parents. Hence, we use the former as the benchmark because it does not require a coresidence restriction and interpret the difference between both sources as indicative of the size of coresidence bias.

We assess the impact of coresidence on these 16 indicators in two dimensions: First, we quantify the average size of the coresidence bias (i.e., the average difference between sources as a percentage of the value computed with retrospective information) for each indicator. Second, we analyze to what extent these indicators provide valuable information to rank economies or cohorts according to the level of intergenerational mobility. We compute the Spearman rank correlation between the IGM indicators using our two data sources to evaluate whether the rankings derived from one of them are consistent with the alternative source.

IV.1 Data and measurement

Data. We use data from two sources that contain information for 18 countries in Latin America: Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Peru, Uruguay, and Venezuela.

First, we use Latinobarometro opinion survey, which has been previously used to document IGM in Latin America (see Neidhöfer et al., 2018). This survey is nationally representative and contains information about the educational attainment of each individual responding the questionnaire plus the information about parents' educational attainment (i.e., each individual is asked about the highest educational attainment of her parents). We include in our sample individuals who were born between 1935-1995 and were at least 23 years old when they answered the survey. For each country, we pool the waves 1998, 2000-2011, 2013 and 2015, and normalize the survey weights over different waves. The data set contains information about educational attainment that can be coded to have years of schooling censored at 15[5] and completed level of education that takes values 1 for iliterate, 2 for incomplete primary, 3 for complete primary, 4 for incomplete secondary or technical, 5 for complete secondary or technical, 6 for incomplete higher education, and 7 for complete higher education.

Second, we use census data obtained from Integrated Public Use Microdata Series-International (IPUMS-International, IPUMS, 2019), which provides samples (typically 10%) of the full-count microdata. The data collection is organized at the household level and is possible to link individuals who live with their parents in the same household at the time of the interview using a variable that details the relationship between each individual and the household head. We use individuals aged 21-25 years linked to their probable father and/or probable mother according to the procedures used by IPUMS for family interrelationships.[6] Table A3 provides the detail of the samples that we use and the availability of educational attainment information. The data set contains a variable with years of schooling (available in a subset of census samples) that we censor at 15 years and a categorical variable (available for all our census samples) that takes values 1 for less than secondary, 2 for primary education, 3 for secondary education and 4 in the case of tertiary education. These levels do not represent any particular country system and are based on a recoding done by IPUMS (2019).[7] We measure educational attainment of parents as the highest attainment among the available parents to be consistent with the information provided by Latinobarometro

[5]This variable is continuous from 0 to 12, and then we code incomplete university or technical training as 13, complete technical training as 14, and complete university as 15

[6]More details can be found in the following link: https://usa.ipums.org/usa/chapter5/chapter5.shtml.

[7]This variable applies, to the extent possible, the United Nations standard of six years of primary schooling, three years of lower secondary schooling, and three years of higher secondary schooling.

opinion survey.

Measurement. We compute 16 indicators of intergenerational mobility in education that can be classified within the concepts of absolute mobility, relative mobility, and movement and have been recently used in the literature. A description of them was provided in Table 1 of section II. For each census sample, we use individuals' respective birth years to identify a sample in Latinobarometro survey that represent the same 5-year birth cohort and country. In total, we are able to identify up to 72 samples, each one a different country and 5-year birth cohort with information available in both data sources.

IV.2 Results

We estimate 16 educational IGM indicators in both data sets and end up with at most 72 country-birth-cohorts that are available in both data sets. Descriptive statistics of the full set of estimates for the census data and Latinobarometro survey can be found in the Appendix (see Table A4 and Table A5, respectively). Using the subset of estimates for country-birth-cohorts that are available in both data sources, we compute the average difference and the Spearman rank correlation (see Table 4).

In terms of the size of coresidence bias (see the column average difference in Table 4), our findings show varying levels of bias going from less than 1% to more than 10%.[8] In the case of absolute and relative mobility, there are indicators with a relatively small bias (for example, UCP and CER050). In contrast, all the indicators of movement here considered have an average bias greater than 10%. In line with the results of Emran and Shilpi (2018), the expected rank for children with parents in the bottom half of the distribution (CER050) show the smallest bias of all the indicators. When comparing the IGRC vs. the IGPC, which was the focus of the previous section, we find a larger average bias in the case of IGRC. However, the bias is positive on average, in contrast to the empirical regularity stated in Emran et al. (2018).

[8] The table reports averages, a visualization of the distribution for each indicator using a boxplot can be found in Figure A2 of the Appendix.

When we assess the level of re-ranking or how aligned are rankings produced with these two sources (see the column rank correlation in Table 4), we find some striking results. First, all the indicators of absolute mobility show relatively high-rank correlations (i.e., the ranking by the level of mobility with one source is close to the ranking with the alternative source). Second, the indicators of relative mobility show varying levels of rank correlation that do not follow the size of the coresidence bias. For example, CER050 has the lowest bias but also one of the lowest rank correlations. In contrast, the IGRC has the highest bias and the highest rank correlation. This suggests that researchers should be careful when pooling IGRC estimates with coresident samples together with those that use all children, but they still could use IGRC to rank economies in terms of relative mobility when all the estimates come from coresident samples. Third, in line with the results for the IGRC, the indicators of movement show relatively large bias and relative high-rank correlation.

Figure 1 provides visual evidence that highlights how some indicators computed using census data are better aligned to those obtained from the social survey Latinobarometro. In this case, measures of absolute mobility such as bottom-upward mobility are close and clearly more spread across the 45-degree line when compared to measures of relative mobility such as IGRC or IGPC. As a consequence of this, we find small bias and high rank correlation in these measures of absolute mobility. A visualization of how the rankings change between sources is provided in Figure A3 of the Appendix. It highlights how some country-birth-cohorts that appear to be highly mobile when using BHQ4 (rank correlation lower than 0.16) with full sample become part of the samples with low level of mobility when using the coresidents (lines crossing from top to bottom in the graph). In contrast, the ranking appear much more stable with BUM-primary (rank correlation 0.91).

So far we have assumed that any difference between the estimates computed with Latinobarometro opinion survey and census data are because of coresidence bias. However, some difference may appear just because of sampling variation too. To put the magnitude of the

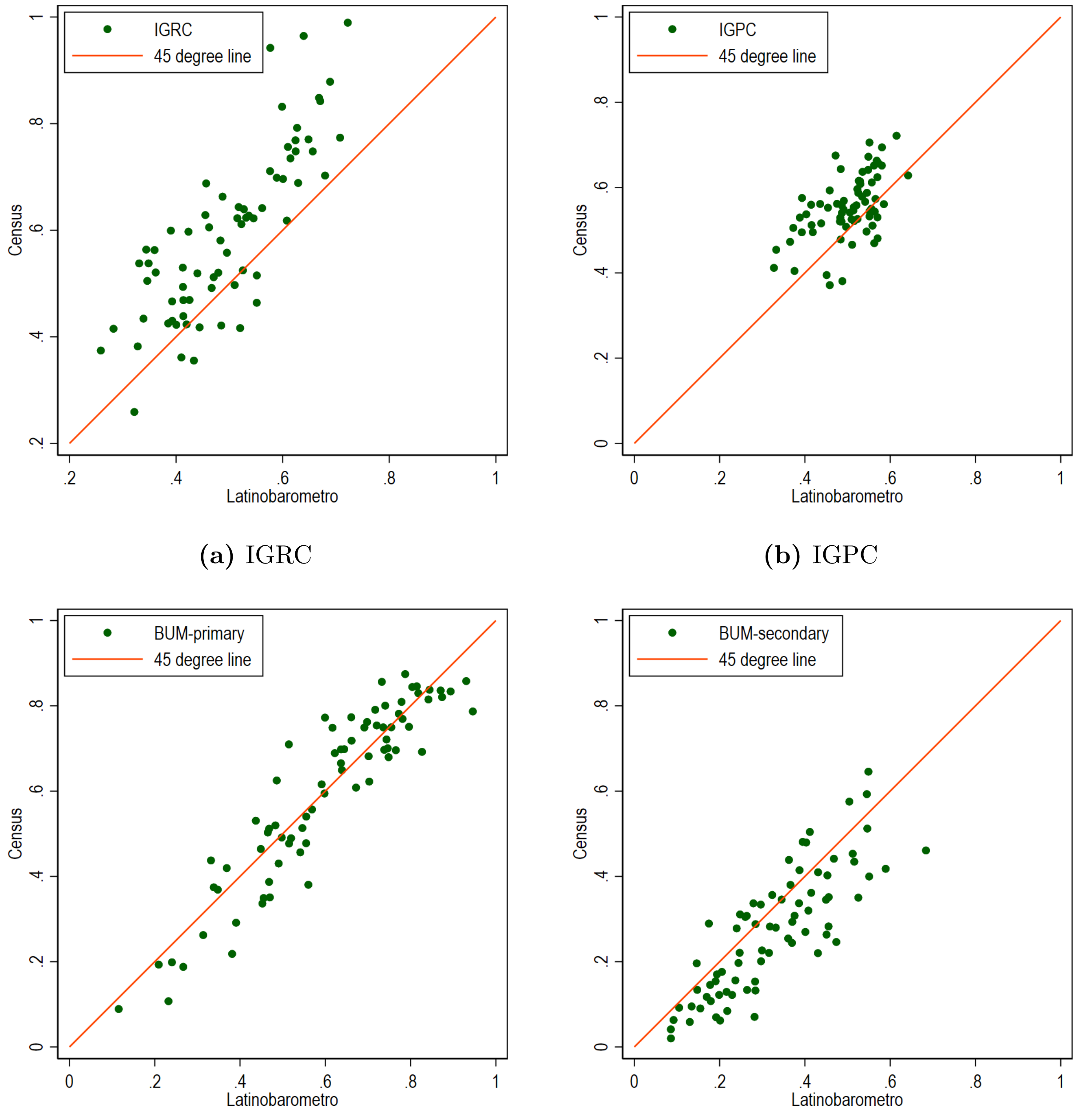
Table 4: Comparison of estimates using a coresident sample (census data) and those with coresidents and non-coresidents (social survey with retrospective information)

Indicator	Average difference (%)	Rank correlation
Absolute mobility		
UCP	0.693	0.551
BUM-primary	-2.199	0.910
YOS	-2.959	0.718
TDM-secondary	12.844	0.551
TDM-primary	14.705	0.737
BUM-secondary	-17.127	0.855
CAT	-30.847	0.744
MIX	-30.951	0.702
Relative mobility		
CER050	6.361	0.186
IGPC	10.854	0.490
IGSC	12.448	0.368
IGRC	18.817	0.820
BHQ4	40.174	0.164
Movement		
M1	-10.812	0.766
M2	-12.159	0.747
DIF	-13.032	0.799

Notes: This table uses estimates of 16 indicators of intergenerational mobility described in Table 1 computed using Latinobarometer social survey and census data. The former contains retrospective information about parents' educational attainment while the latter uses a sample of coresidents. The first column reports the average difference between the estimates of both sources as a percentage of the indicator computed with the former. The second column reports for each indicator the Spearman rank correlation coefficient relating the estimates using one source to the estimates using the alternative source.

bias and re-ranking in context, we also run a similar analysis that compares some of the IGM measures computed with two different data sources that contain retrospective information for 9 countries (Brazil, Chile, Colombia, Ecuador, Guatemala, Mexico, Nicaragua, Panama, and Peru). Table A7 in the Appendix shows the rank correlation and average differences of the set of IGM measures computed with Latinobarometro and nationally representative

Figure 1: IGM with retrospective information vs. coresident samples



(a) IGRC

(b) IGPC

(c) BUM-primary

(d) BUM-secondary

Notes: The figure shows estimates for up to 72 samples (each one a different country and 5-year birth cohort) of 4 indicators of intergenerational mobility as described in Table 1. They are computed with a social survey that contains retrospective information (Latinobarometro) and a coresident sample from census data using individuals aged 21-25 years.

household surveys[9] and made available in Neidhöfer et al. (2018).[10] In terms of average differences, we find average differences of more than 5% in relative mobility computed with the IGRC and IGPC but around 4% with IGSC. In contrast, indicators of absolute mobility and movement show smaller differences. In the case of rank correlations, the bottom-upward mobility indicator show the highest alignment while the IGPC and IGSC provide very small rank correlation. This suggests that even in the absence of coresidence bias, some indicators of relative mobility are not very reliable to rank economies by the level of IGM. In their analysis, Neidhöfer et al. (2018) omit cohorts with less than 200 observations when analyzing trends over time. When we apply the same constraint, our main findings still hold, which suggests that the differences are not driven by this set of estimates computed with very small samples that are arguably less reliable.

V Conclusion

Researchers and journal editors are cautious about using coresident samples to estimate intergenerational mobility indicators because of potential sample selection bias from truncation. However, there is scarce empirical evidence on how sensitive these measures are to coresidence restriction (i.e., estimating an indicator using only individuals living with their parents).

This paper contributes to the understanding of the impact of coresidence bias on educational IGM. We begin re-examining a recent conclusion that the intergenerational correlation is less affected by coresidence bias than the intergenerational regression. We find that the conclusion depends on the setting in which researchers are estimating educational mobility: if both, the variance of years of schooling of parents and the variance of years of schooling of children are truncated, then the result is not warranted. We also show that a comparison of estimates pooling a large number of birth cohorts with a full sample against those with

[9] Table A6 in the Appendix specifies what household surveys and waves are being used.

[10] Figure A4 in the Appendix shows scatter plots of these comparisons.

coresidents sample, tends to favor (in terms of bias) the indicator that varies less across birth cohorts (usually the correlation coefficient).

Furthermore, we take advantage of two data sources to investigate how coresidence bias affects different measures of intergenerational mobility in education for a large number of countries and birth cohorts. Our main empirical findings are threefold: First, some indicators of absolute mobility computed with coresident samples provide meaningful information to rank economies by the level of mobility and show low coresidence bias levels. Second, the Pearson correlation coefficient is usually insufficient to rank economies across time and space despite having lower bias than alternative indicators of relative mobility. Third, the Pearson correlation coefficient gives a low-rank correlation even when comparing two sources of information where none suffers from the coresidence restriction. Similarly, the rank-based mobility indicators produce significant levels of re-ranking even when coresidence bias is not an issue.

The fact that some indicators offer relatively high coresidence bias together with high-rank correlation implies that researchers can use these indicators with confidence to rank economies and/or cohorts using the same metric estimated with coresident samples. However, researchers should be careful when comparing the same indicator computed with coresident samples versus full samples.

Our work underlines that census data is a viable alternative for further research on intergenerational mobility in education. It opens research opportunities in economies that lack alternative data and offers historical options in places with good data today but not in the past. The fact that census data can be used to study IGM at a disaggregated geographical level also opens up possibilities to find credible natural experiments to shed some light on the drivers of IGM in education.

References

Alesina, A., Hohmann, S., Michalopoulos, S., & Papaioannou, E. (2020). Religion and Educational Mobility in Africa. *NBER Working Paper*.

Alesina, A., Hohmann, S., Michalopoulos, S., & Papaioannou, E. (2021). Intergenerational Mobility in Africa. *Econometrica*, *89*(1), 1–35.

Asher, S., Novosad, P., & Rafkin, C. (2021). Intergenerational Mobility in India: Estimates from New Methods and Administrative Data. *Working Paper. Available at: http://paulnovosad.com/pdf/anr-india-mobility.pdf (accessed November 2021))*.

Card, D., Domnisoru, C., & Taylor, L. (2018). The Intergenerational Transmission of Human Capital: Evidence from the Golden Age of Upward Mobility. *NBER Working Paper*(25000), 1–69.

Derenoncourt, E. (2021). Can You Move to Opportunity? Evidence from the Great Migration. *Forthcoming, American Economic Review*.

Deutscher, N., & Mazumder, B. (2021). Measuring Intergenerational Income Mobility: A Synthesis of Approaches. *FRB of Chicago Working Paper*, *09*.

Dodin, M., Findeisen, S., Henkel, L., Sachs, D., & Schüle, P. (2021). Social Mobility in Germany. *CEPR Discussion Paper*, *01*.

Emran, M. S., Greene, W., & Shilpi, F. (2018). When Measure Matters: Coresidency, Truncation Bias, and Intergenerational Mobility in Developing Countries. *Journal of Human Resources*, *53*(3), 579–607.

Emran, M. S., & Shilpi, F. (2018). Estimating Intergenerational Mobility with Incomplete Data: Coresidency and Truncation Bias in Rank-Based Relative and Absolute Mobility Measures. *Policy Research Working Paper*, *8414*(April).

Emran, M. S., & Shilpi, F. (2019). Economic Approach to Intergenerational Mobility: Measures, Methods, and Challenges in Developing Countries. *WIDER Working Paper*(98).

Feigenbaum, J. J. (2018). Multiple Measures of Historical Intergenerational Mobility: Iowa 1915 to 1940. *Economic Journal*, *128*(612), F446–F481.

Fields, G. S., & Ok, E. A. (1996). The Meaning and Measurement of Income Mobility. *Journal of Economic Theory*, *71*(2), 349–377.

Francesconi, M., & Nicoletti, C. (2006). Intergenerational Mobility and Sample Election in Short Panels. *Journal of Applied Econometrics*, *21*, 1265–1293.

Geng, Y. (2020). Intergenerational Mobility in China across Space and Time. *mimeo*.

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

Hilger, N. (2016). The Great Escape: Intergenerational Mobility in the United States Since 1940. *NBER Working Paper*(21217).

IPUMS. (2019). *Integrated Public Use Microdata Series, International: Version 7.2 [dataset].* Minneapolis, MN: IPUMS: Minnesota Population Center.

Munoz, E. (2021a). The Geography of Intergenerational Mobility in Latin America and the Caribbean. *Stone Center Working Paper*, *29*.

Munoz, E. (2021b). Intergenerational Educational Mobility within Chile. *mimeo*.

Narayan, A., Van der Weide, R., Cojocaru, A., Lakner, C., Redaelli, S., Gerszon Mahler, D., ... Thewissen, S. (2018). *Fair Progress?: Economic Mobility Across Generations Around the World.* The World Bank.

Neidhöfer, G., Serrano, J., & Gasparini, L. (2018). Educational Inequality and Intergenerational Mobility in Latin America: A New Database. *Journal of Development Economics*, *134*, 329–349.

Torche, F. (2019). Educational Mobility in Developing Countries. *WIDER Working Paper*(88), 1–31.

Van der Weide, R., Ferreira de Souza, P., & Barbosa, R. (2020). Intergenerational Mobility in Education in Brazil. *mimeo*.

Van der Weide, R., Lakner, C., Gerszon Mahler, D., Narayan, A., & Ramasubbaiah, R. (2021). Intergenerational Mobility Around the World. *mimeo*.

Appendices

The appendix provides additional tables and figures, and other relevant information.

Table A1 presents evidence of coresidence bias by cohort for Guatemala.

Table A2 presents evidence of coresidence bias by cohort for Eduador.

Table A3 lists the countries and census years used in this study and the availability of information about years of education and/or education categories.

Table A4 provides summary statistics of indicators computed with census data.

Table A5 provides summary statistics of indicators computed with Latinobarometro social survey.

Table A6 lists the countries and respective household surveys from which intergenerational mobility indicators are derived in Neidhöfer et al. (2018).

Table A7 reports the comparison of intergenerational mobility in education with different data sources (i.e., Latinobarometro social survey versus nationally representative household surveys).

Figure A1 displays estimates of intergenerational mobility across cohorts in India.

Figure A2 provides information about the distribution of the difference between estimates of IGM (see detail of indicators in Table 1) using coresident samples (census data) and full samples (Latinobarometro social survey) for up to 72 country-birth-cohorts.

Figure A3 compares the way in which different country-cohorts are ranked with two different sources of information according to their level of IGM computed with the IGPC and BUM.

Figure A4 displays scatter plots of 11 indicators of intergenerational mobility in education estimated with coresident samples obtained from census data against those using retrospective information with Latinobarometro social survey.

Figure A5 compares the way in which different country-cohorts are ranked with two different sources of information according to their level of IGM computed with the IGPC and BUM.

Table A1: Coresidence bias and relative mobility in Guatemala's ENCOVI household survey

	Age groups (children)					
	21-25	26-35	36-45	46-55	56-65	21-65
IGRC	.57	.69	.76	.83	.87	.75
IGPC	.56	.59	.59	.58	.64	.61
IGRC (coresident sample)	.55	.66	.72	.68	.7	.63
IGPC (coresident sample)	.56	.61	.64	.55	.57	.59
Bias in IGRC (%)	-3.4	-4.1	-5.1	-17	-20	-16
Bias in IGPC (%)	1.2	3.6	8.9	-6.2	-11	-2
Ratio of SD (σ_p/σ_c)	.97	.85	.78	.71	.74	1.2
Ratio of SD (coresident sample)	1	.92	.89	.8	.81	1.1
Bias in ratio of SD (%)	4.8	8	15	14	10	17
N	4934	7206	5291	3958	2721	24110
Coresidence rate (%)	57	29	14	6.9	3.2	25

Notes: The table report estimates of the intergenerational regression coefficient (IGRC) and the intergenerational Pearson correlation coefficient (IGPC) computed for different age groups and pooling all these groups. These estimates use years of schooling (YOS) censored at 15 for children and their parents, and for the latter use the highest level when information about both parents is available, and the one available when that is not the case. We use all the children of a given age and then restrict the sample to only those that coreside with at least one parent (i.e., coresident sample). We report the bias in these indicators as percentage of the value computed with the full sample (coresidents and no coresidents). We report the standard deviation (SD) of YOS of parents (σ_p) over the SD in YOS of children (σ_c), and the bias computed as percentage of the ratio estimated with the full sample. The row N reports the number of children used in the estimation with the full sample, and the coresidence rate in the last row indicates the percentage of all children (i.e., N) living with at least one parent.

Table A2: Coresidence bias and relative mobility in Ecuador's ECV household survey

	Age groups (children)					
	21-25	26-35	36-45	46-55	56-65	21-65
IGRC	.4	.49	.54	.62	.7	.56
IGPC	.48	.53	.54	.54	.59	.56
IGRC (coresident sample)	.39	.48	.54	.66	.66	.48
IGPC (coresident sample)	.48	.52	.53	.58	.55	.53
Bias in IGRC (%)	-1.6	-1.9	-.11	6.6	-5.6	-14
Bias in IGPC (%)	.32	-.44	-1.8	6.8	-7.4	-5
Ratio of SD (σ_p/σ_c)	1.2	1.1	1	.87	.84	1
Ratio of SD (coresident sample)	1.2	1.1	.98	.87	.83	.91
Bias in ratio of SD (%)	2	1.5	-1.7	.19	-1.8	9.9
N	8095	14929	12296	9440	6555	51315
Coresidence rate (%)	50	24	12	8.4	4.7	20

Notes: The table report estimates of the intergenerational regression coefficient (IGRC) and the intergenerational Pearson correlation coefficient (IGPC) computed for different age groups and pooling all these groups. These estimates use years of schooling (YOS) censored at 15 for children and their parents, and for the latter use the highest level when information about both parents is available, and the one available when that is not the case. We use all the children of a given age and then restrict the sample to only those that coreside with at least one parent (i.e., coresident sample). We report the bias in these indicators as percentage of the value computed with the full sample (coresidents and no coresidents). We report the standard deviation (SD) of YOS of parents (σ_p) over the SD in YOS of children (σ_c), and the bias computed as percentage of the ratio estimated with the full sample. The row N reports the number of children used in the estimation with the full sample, and the coresidence rate in the last row indicates the percentage of all children (i.e., N) living with at least one parent.

Table A3: Census data sets and availability of information about education

Country	Census years	Years of schooling	Categories
Argentina	1970, 1980, 1991, 2001	Yes	Yes
Bolivia	1976, 1992, 2001, 2012	Yes	Yes
Brazil	1960, 1970, 1980, 1991, 2000, 2010	Yes, except 2010	Yes
Chile	1970, 1982, 1992, 2002	Yes	Yes
Colombia	1973, 1985, 1993, 2005	Yes, except 1993 censored	Yes
Costa Rica	1973, 1984, 2000, 2011	Yes	Yes
Dominican Republic	1981, 2002, 2010	Yes	Yes
Ecuador	1974, 1982, 1990, 2001, 2010	Yes	Yes
El Salvador	1992, 2007	Yes	Yes
Guatemala	1964, 1973, 1981, 1994, 2002	Yes	Yes
Honduras	1974, 1988, 2001	Yes	Yes
Mexico	1970, 1990, 1995, 2000, 2010, 2015	Yes	Yes
Nicaragua	1971, 1995, 2005	Yes	Yes
Panama	1960, 1970, 1980, 1990, 2000, 2010	Yes	Yes
Paraguay	1962, 1972, 1982, 1992, 2002	Yes	Yes
Peru	1993, 2007	No, censored	Yes
Uruguay	1963, 1975, 1985, 1996, 2006, 2011	Yes, except 2011	Yes
Venezuela	1971, 1981, 1990, 2001	Yes	Yes

Notes: The categorical educational variable is coded with values 1-4 as: less than primary completed, primary completed, secondary completed, and university completed. Some census samples available in the original source where there is information about education but the data is not organized in households are excluded because we cannot link individuals to their parents.

Table A4: Summary statistics of indicators computed with census data

	Mean	Std. dev.	Min	Max	N
YOS	0.69	0.10	0.35	0.83	71
CAT	0.45	0.12	0.10	0.67	76
MIX	0.44	0.12	0.10	0.65	76
BUM-primary	0.60	0.21	0.09	0.87	76
BUM-secondary	0.27	0.15	0.02	0.65	76
UCP	0.76	0.11	0.41	0.93	76
TDM-primary	0.08	0.04	0.02	0.22	76
TDM-secondary	0.24	0.11	0.07	0.59	76
IGRC	0.59	0.16	0.26	0.99	71
IGPC	0.55	0.08	0.37	0.72	71
IGSC	0.56	0.07	0.38	0.68	71
CER050	36.43	2.12	31.68	42.28	71
BHQ4	0.10	0.03	0.03	0.18	71
M1	3.47	0.74	1.34	4.92	71
M2	2.65	0.77	0.91	4.14	71
DIF	2.85	0.82	0.92	4.38	71
Observations	76				

Notes: The columns reports the mean, the standard desviation, the minimun and the maximum values for the indicators calculated using census data.

Table A5: Summary statistics of indicators computed with Latinobarometro

	Mean	Std. dev.	Min	Max	N
YOS	0.71	0.10	0.37	0.90	1026
CAT	0.64	0.08	0.34	0.82	1026
MIX	0.63	0.08	0.34	0.77	1026
BUM-primary	0.58	0.20	0.12	1.00	1026
BUM-secondary	0.31	0.15	0.05	0.82	1026
UCP	0.76	0.11	0.33	0.96	1025
IGRC	0.50	0.12	0.15	0.83	1026
IGPC	0.50	0.07	0.27	0.71	1026
IGSC	0.49	0.07	0.24	0.69	1026
CER050	36.92	3.51	22.22	46.35	1026
BHQ4	0.12	0.03	0.04	0.22	1026
M1	3.88	0.61	2.22	5.26	1026
M2	3.02	0.70	0.92	4.73	1026
DIF	3.29	0.73	1.05	5.38	1026
Observations	1026				

Notes: The columns reports the mean, the standard desviation, the minimun and the maximum values for the indicators calculated using Latinobarometro.

Table A6: Nationally representative household surveys

Country	**Name of survey**	**Acronym**	**Survey waves**
Brazil	*Pesquisa Nacional por Amostra de Domicilios*	PNAD	2014
Chile	*Encuesta de Caracterización Socioeconómica Nacional*	CASEN	2006, 2009, 2011, 2013, 2015
Colombia	*Encuesta Nacional de Condiciones de Vida*	ECV	2003, 2008, 2010-2013
Ecuador	*Encuesta de Condiciones de Vida*	ECV	1994, 1995, 1998, 2006
Guatemala	*Encuesta Nacional sobre Condiciones de Vida*	ENCOVI	2000, 2006, 2011
México	*Encuesta Nacional sobre Niveles de Vida de los Hogares*	MXFLS	2002, 2005-2006, 2009-2012
Nicaragua	*Encuesta Nacional de Hogares sobre Medición de Nivel de Vida*	EMNV	1998
Panama	*Encuesta de Niveles de Vida*	ENV	1997, 2003, 2008
Peru	*Encuesta Nacional de Hogares*	ENAHO	2001-2015

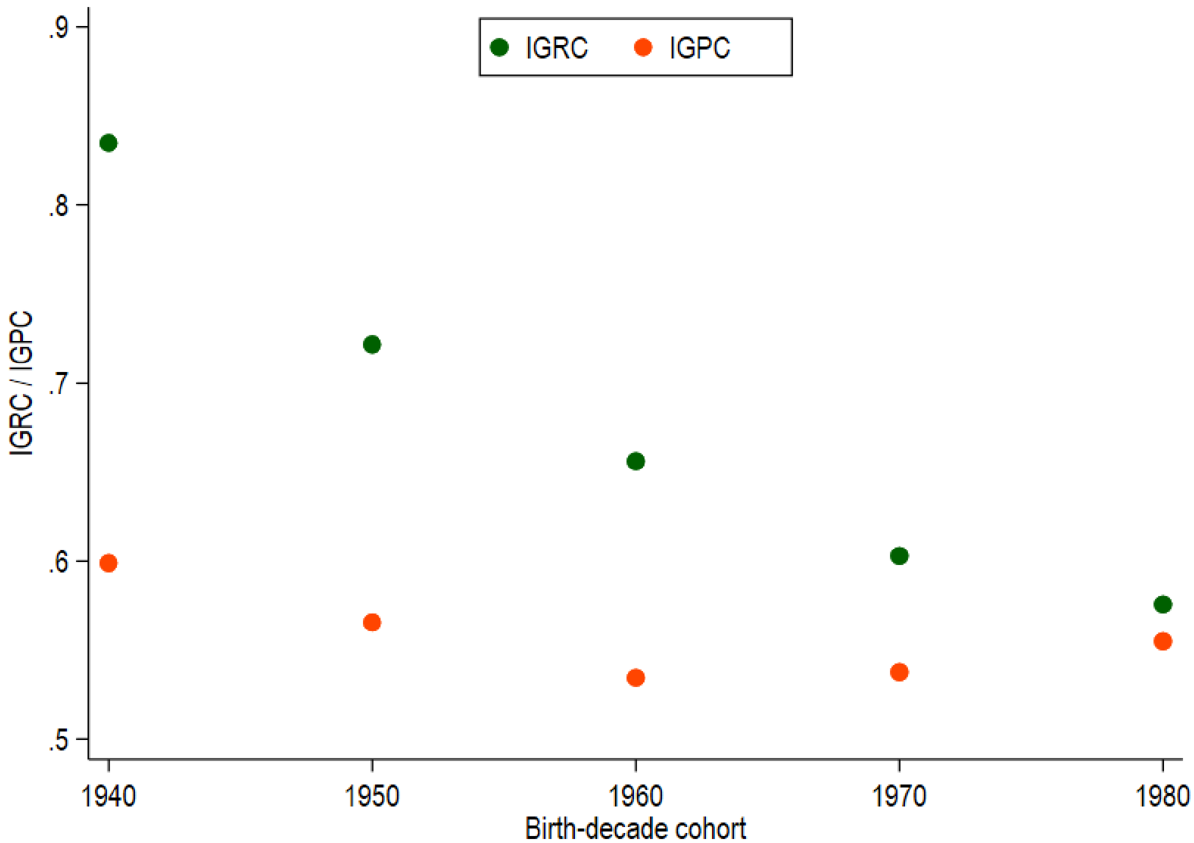
Notes: Nationally representative household surveys used to compute intergenerational mobility estimates in Neidhöfer et al. (2018).

Table A7: Comparison of indicators with retrospective information but different data sources (social surveys vs. household surveys)

Indicator	Average difference (%)	Rank correlation
Absolute mobility		
BUM-secondary	-1.985	0.840
UCP	3.639	0.518
Relative mobility		
IGSC	3.642	0.067
IGPC	7.019	0.050
IGRC	13.210	0.699
Movement		
M2	-0.438	0.590
M1	-0.961	0.638

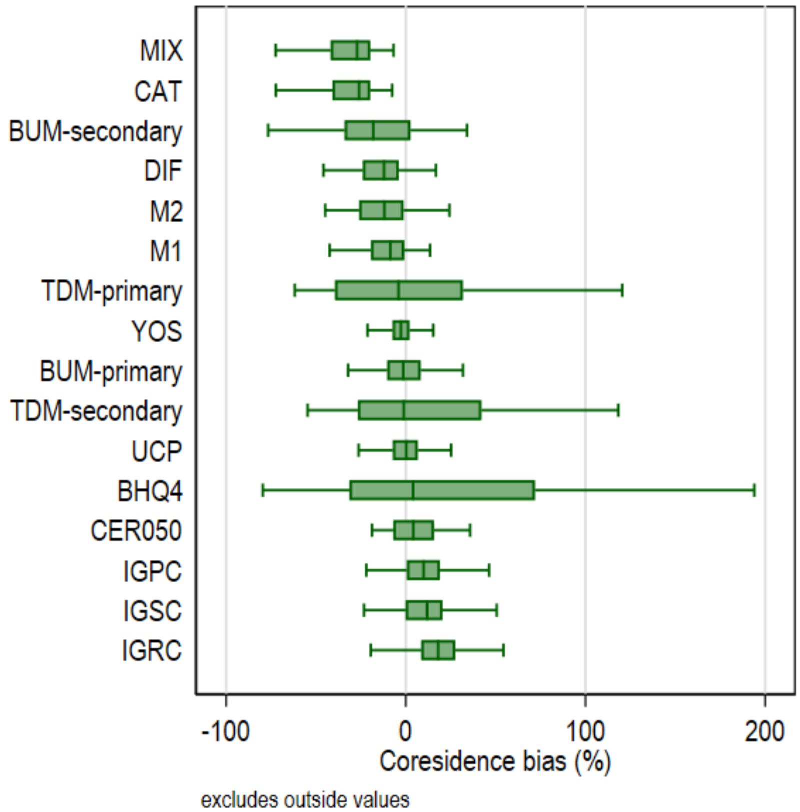
Notes: The first column reports the average difference as percentage of the indicator computed using Latinobarometro. The second column reports the Spearman rank correlation coefficient for 7 indicators of intergenerational mobility described in Table 1 computed using Latinobarometro social survey and other alternative nationally representative household surveys (see details in Table A6). The sample include multiple cohorts for 9 countries that sum up to 113 estimates. The source of these estimates is Neidhöfer et al. (2018).

Figure A1: Intergenerational mobility across birth cohorts in India



Notes: The figure display estimates of the intergenerational Pearson correlation coefficient and the intergenerational regression coefficient by birth-decade cohort in India (1940=1940/1949, 1950=1950-1959, 1960=1960-1969, 1970=1970-1979, and 1980=1980-1989). The source of these estimates is Narayan et al. (2018).

Figure A2: Boxplot of differences between coresident sample and full samples



Notes: The figure provides information about the distribution of the difference between estimates of IGM (see detail of indicators in Table 1) using coresident samples (census data) and full samples (Latinobarometro social survey) for up to 72 country-birth-cohorts. The difference is reported as percentage of the estimate with the full sample.

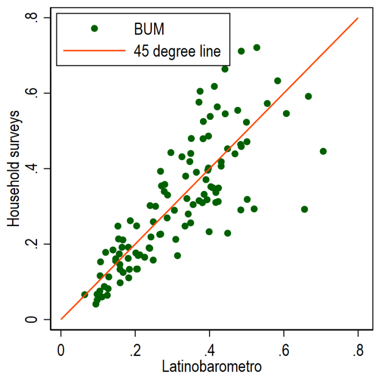
Figure A3: Comparison of rankings with full sample and coresident sample

(a) BUM-primary

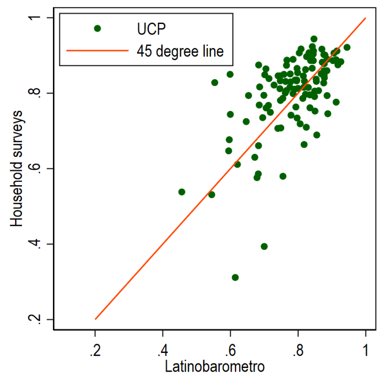
(b) BHQ4

Notes: The figure plots lines connecting the rank of estimates for the same country-cohorts computed with two data sources (social survey vs. census data, the former with retrospective information and the latter being a coresident sample). It is sorted according to the rank computed using Latinobarometro social survey. The sample includes multiple 5-year birth cohorts for 16 countries that sum up to 72 estimates.

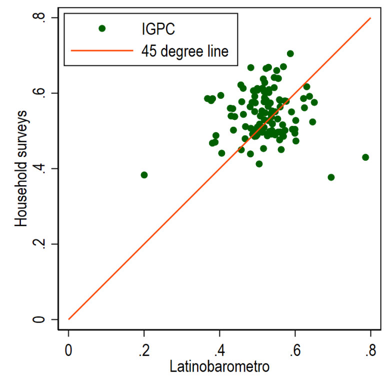
Figure A4: Comparison of indicators with retrospective information but different data sources



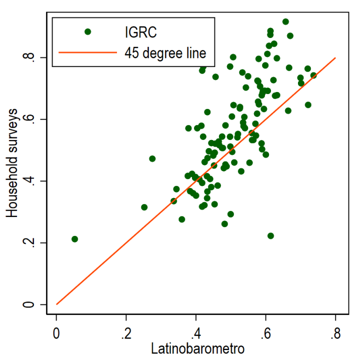
(a) BUM



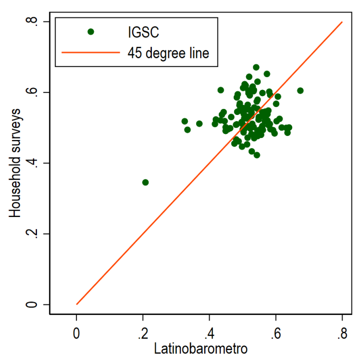
(b) UCP



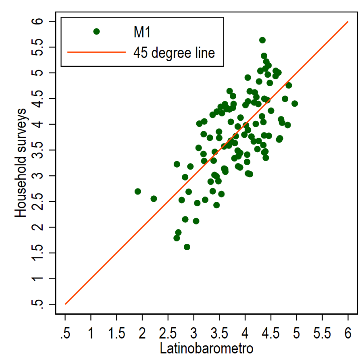
(c) IGPC



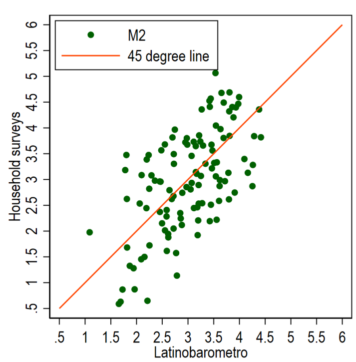
(d) IGRC



(e) IGSC



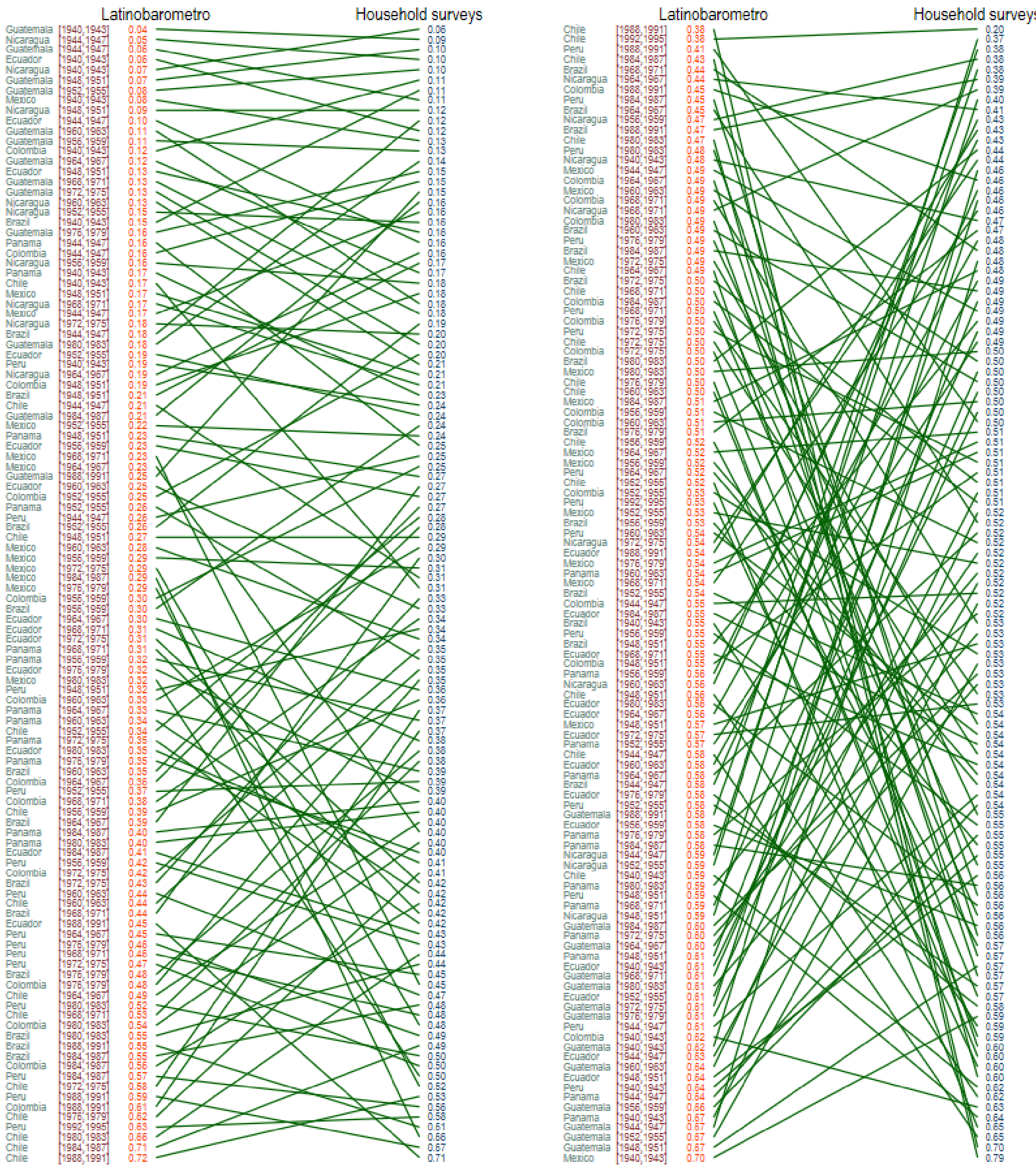
(f) M1



(g) M2

Notes: The figure plots estimates for the same country-cohorts computed with two data sources (social survey vs. household survey, both with retrospective information). The sample include multiple cohorts for 9 countries that sum up to 113 estimates. The source of these estimates is Neidhöfer et al. (2018).

Figure A5: Comparison of rankings with different data sources



(a) BUM

(b) IGPC

Notes: The figure plots lines connecting the rank of estimates for the same country-cohorts computed with two data sources (social survey vs. household survey, both with retrospective information). It is sorted according to the rank using Latinobarometro social survey. The sample include multiple cohorts for 9 countries that sum up to 113 estimates. The source of these estimates is Neidhöfer et al. (2018).