Household Recombination, Retrospective Evaluation, and Educational Mobility over 40 Years

Andrew Foster and Sveta Milusheva

December 29, 2015

Abstract

Despite longstanding belief in certain circles that investment in primary health care and education can help to encourage reductions in inequality and increases in intergenerational economic mobility, evidence is scarce due to the lack of systematically collected data from developing countries that links households over multiple decades. Bangladesh would seem an especially fruitful avenue for looking at these issues given international recognition of its success in improving basic health care. In this paper we use a newly collected survey data connected to the Matlab Demographic Surveillance System (DSS) maintained by the International Center for Diarrhoeal Disease Research in Bangladesh (ICDDRB) to take a first look at this issue. A novel insight from this paper is that standard methods for correcting sampling weights in panel data do not adequately account for the process of household formation and dissolution.We develop a new approach to weighting that requires the kind of information available in the context of a DSS, and use these weights to look at long term changes in educational investment of households in the Matlab area. We show that a substantial rise in average educational investment among children 6-16 has been accompanied by high levels of economic mobility but little reduction in economic inequality.

1 Introduction

While the process of economic mobility is often studied by examining the long term prospects of individuals with different background characteristics, in practice changes in the well-being of individuals are importantly determined by the households and families in which these individuals are embedded. This relationship is perhaps most clear with respect to household level measures such as expenditure or poverty, but even arises in the study of mobility with respect to something as individualized as child investment such as in education. Consider the question, for example, of whether educational inequality at one point in time persists into subsequent periods and the degree to which this persistent inequality is influenced by various types of public programs and services. One cannot simply look at educational change for a given person over twenty years. Similarly one cannot learn much by comparing the education of a child of one mother to the education of another child born to the same mother twenty years later. Childbearing and education are intrinsically tied to the span of childbearing of the mother and to the particular ages of the children. Following biological lines of descent, that is comparing education of a child to her mother and grandmother has a certain logic, but given the distributions in age at childbearing, this can lead to substantial selectivity if education or other measures of child investment are only measured at discrete intervals of 10-20 years. From this perspective, and accounting for the role of households in determining child investment, it would seem advantageous to look at educational mobility by comparing the education of children of a particular age at one point in time in a particular household to the education of children of that same age some years later in another household that is related in some well-specified way to the previous household.

Analysis of mobility in this way might be relatively straightforward using long-term household panel survey data if households stayed fixed over time; however, they do not. A household at one point in time morphs through a process of household division and fusion, which we combine under the rubric of household recombination. Not only is it necessary to account for this process of recombination in the evaluation of data on economic mobility but the process of recombination in turn can affect the process of economic mobility. Depending on the nature of the sampling processes and how recombination proceeds over time it may be necessary to reweight data based on weights that may themselves be endogenous with respect to household recombination.

The importance of understanding household formation as an element of analyses of other development outcomes has been previously recognized in the literature. Foster 1993, for example, noted "It is increasingly recognized that certain demographic variables should be treated as endogenous in analyses of economic and demographic data from developing countries, very little is known about the implications of the fact that joint residence is itself a choice variable." In subsequent work looking at the relationship between household division and inequality, Foster and Rosenzweig 2002 argue "An improved understanding of the determinants of household division is thus useful not only for dealing with the potential selectivity of panel designs that drop dividing households, but in studying household behaviour and income change generally." But both papers focus only on the process of household division and, in part as a consequence of this, neither deals explicitly with the issue of how sampling composition is affected by the process of household recombination and thus influences the construction of sample weights.

The process of using weights to adjust for sample attrition has, of course, achieved substantial attention in the literature (Fitzgerald, Gottschalk, and Moffitt 1998, Moffit, Fitzgerald, and Gottschalk 1999). Generally, one inflates the weight of observed sampling unit households based on the assumption that attrition is random with respect to processes of interest in the data conditional on the observables. Of course, this assumption may not be correct in general and even if it is, the outcomes of the attritting population may be sufficiently distinct from the observed population (e.g., attrition through mortality) that one cannot sensibly combine the outcomes of the observed and attritted population into a single metric. But little attention has been given to the question of weighting in a setting in which the problem is not sample attrition but the process of household recombination. In this case observability may not be as much of an issue as in the case of attrition, but the flow of people across households and the fact that the sampling unit for a survey is generally the household creates a new set of problems.

To understand this point, consider a random sample of households collected at time t and assume that larger (at time t) households are more likely to divide. If all descendant households are followed, then one will correctly measure at time t+1 the distribution of t+1 attributes such as household size. However, if these t+1 households are used to retrospectively construct the mean household size at time t, then the estimate will be too large because large t households are overrepresented in the resulting sample of t+1 households. A similar bias would arise if household size at time t were estimated from a random sample of households

in period t+1. A simple correction in each case would be to inverse weight each household by the number, if available, of co-descendant households with the same antecedent household.

The situation is further complicated in the presence of household fusion. The set of all descendant households in t+1 generated from an initial sample in period t will no longer yield an unbiased estimate of household measures at time t+1 because t+1 households composed of members of multiple antecedent households will be more likely to be selected than would be the case if t+1 households were selected randomly. To correctly constitute a representative sample, as discussed below, one would need to know the sampling probabilities and descent paths of households that were not sampled at t.

This is not just a curiosity. Household surveys are at times used to evaluate the consequences of interventions that were introduced at a previous period, and in some cases retrospective or previously collected data are incorporated into the analysis in order to estimate differential change over time. In such cases we may ask if it is possible to mimic the results of a treatment/comparison design in which a baseline is collected from a random sample at a particular point in time, a set of treatments is assigned to the participants, and then outcomes are evaluated at some endline. Our answer is a tentative yes, but as our application suggests, the data requirements for doing so are extremely demanding.

In this paper, we tackle these larger questions of household recombination, sample selection, and weighting mechanisms in panel datasets by looking at a specific example. We focus, in particular, on the process of educational mobility in the Matlab study area of rural Bangladesh. This area is an ideal setting for three reasons. First, there is a long standing debate about the role of broad-scale public health services as mechanism to increase economic mobility and reduce inequality. Much of this argument has been carried out with reference to countries such as Sri Lanka, Cuba, or Costa Rica or states like Kerala with particular cultural and political conditions that might lead to a predisposition toward relatively equal growth. Bangladesh, by contrast, is seen as something of anomaly as a place that remains poor and in which redistributive forces are not strong, but there has been a suprisingly strong and successful support of primary health care. Indeed a recent issue of the Lancet devoted significant attention to Bangladesh as a kind of paradox in this regard (e.g., AMR et al. 2013). Bangladesh's unusual status has also been recognized by economists such as Sen 2013, in comparison to India, and

by Pitt, R, and Hassan 2012 in terms of the rapid growth in female schooling. Thus Bangladesh seems a particularly interesting case to look at questions of long-term economic mobility in an environment with high quality primary health care.

Second, Matlab in particular was a leader in the development of strategies for advancing primary health care and reproductive health. Through the research of the International Centre for Diarrhoeal Desease Research in Bangladesh (ICDDRB) it was involved in the early evaluation, testing, and delivery of oral rehydration therapy. Moverover, it is perhaps best known in social science circles for its introduction in 1978 of a treatment/comparison design to examine the effects of client-centered family planning and maternal child health services. Treatment and comparison differences in education and other outcomes in 1996 have been studied elsewhere (Roy and Foster 1996,Joshi and Schultz 2007, Schultz 2009). Joshi and Schultz 2007, and the 2012 analysis of the treatment/comparison difference is not the focus of this paper. For present purposes of primary importance is the fact that both treatment and comparison areas were receiving high quality primary health and reproductive services by the early 1990s.

Third, as noted, the data requirements for looking at economic mobility over multiple generations in the presence of changing patterns of household coresidence are quite high, even in the presence of panel data. Indeed, as we show below, what is needed formally to address this issue is precisely what is provided by a typical demographic surveillance system. In particular one needs to be able to link individuals across time through the use of a unique id and one needs to be able to identify coresidence patterns at each point in time. The Matlab study area has been under a process of continuous vital registration since the late 1960s, with periodic censuses. There also have been two comprehensive socio-economic surveys, one in 1996 and one in 2012, which can provide more detailed data on economic circumstances than are available from vital registration data.

We focus specifically on the economic mobility as measured by the education attainment of children 6-16 in 1974, 1996, and 2012 among households that are related by coresidence and are resident in the Matlab area during the corresponding survey. In particular we examine educational investment among children in the previous surveys conditional on educational investment in the 1974 households. We also contrast villages with relatively high and relatively low levels of adult education in 1974. Our findings suggest that while there has been a relatively large expansion in educational investment over this 38 year period, the overall

inequality in educational investment has changed little. On the other hand, consistent with views about the role of primary care in promoting mobility, we see a high degree of economic mobility in the Matlab area over the study period. A two standard deviation difference in educational investment in 1974 is associated with a .7 standard deviation difference in 1996 and a .2 standard deviation difference in 2012. Moreover, while 1974 differences in average adult education were still evident in 1996, they had essentially disappeared by 2012. The convergence seems in part to be related to a fairly uniform decline in household size, slower per capita consumption growth in higher education households, and mixing between high and low-education villages. Migration was higher in the higher education households but lower, conditional on household education, in the high education villages.

In what follows we first dicuss the various data sets from Matlab, Bangldesh that are used in the analysis. In Section 3 we describe in more detail the challenge of constructing cross-sectional measures of the population distribution of outcomes in a given geographic region using household panel data. We then develop corrective sample weights and apply them to the Matlab data. The weighted estimates are then compared to estimates that do not properly account for recombination. In Section 4 these procedures are then used to evaluate the distribution of child schooling investment in Matlab in 1974, 1996 and 2012. We then turn to the question of economic mobility, which raises analogous but distinct problems of estimation. The formal methodology and an analyisis of the circumstances in which it is mostly likely to matter appears in Section 5. Section 6 applies the approach to an analysis of educational mobility from 1974 to 1996 and from 1974 to 2012. After exploring changes in household structure that likely contributed to patterns if inequality change and to educational mobility in Section 7, we conclude.

2 Data

The ICDDR,B in the Matlab region of Bangladesh began to maintain a Health and Demographic Surveillance System registering all births, deaths and migrations starting in 1966. There are data available for the full period on 149 villages, which include over 200,000 people. In 1974 the ICDDR,B conducted the first comprehensive census of the region. Censuses were again conducted in 1982, 1993 and 1996. From this census data we have information on every single household in the region including basic demographic information and some information on

assets.

In addition to the censuses that were collected, in 1996 a Health and Socio-economic Survey (MHSS) was conducted in Matlab. This survey collected detailed economic and social data on a sample of the population, which had not been done before. The goal was to use this data in order to look at the effect of the maternal and child health and family planning intervention on a wide range of outcomes and over a long period (Rahman et al. 1999).

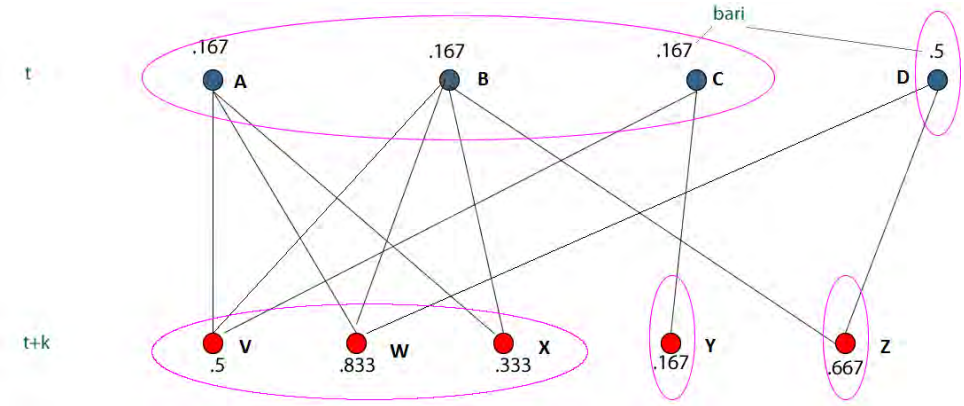
The MHSS sample was selected from the population in 1993. In 1993, there were 38,489 households split among 7,440 baris, which are clusters of households in close physical proximity that are usually linked in a kin network. Of those, 2,883 baris were randomly picked to be part of the sample. Within each bari, one household was randomly selected for the detailed interview. Sampling was done at the bari level because it provides a better representation of family networks as compared to sampling households. There were 102 baris that no longer existed in 1996, and therefore the final number of baris sampled was 2,781. A second household in each bari was also interviewed, but this was not done randomly, so most researchers conduct analyses using only the first household, and we also focus on this primary sample.

A follow-up socieoeconomic survey, MHSS2, was collected starting in 2012. Households were sampled if they included someone who was part of the 1993 sample (zero order sampling link), who was the child or spouse of such an individual and was coresident with that individual after 1993 (first order sampling link), and if they were the spouse or child and coresident with a first order link. A subset of households that did not have any antecedents (defined through this process of links) in 1993 was also included in the sample. Migrants were also studied, but analysis of those individuals is not included in this paper. There were 24,795 individuals and 5952 households included in the 2012 sample.

3 Creating Cross-sectional Weights from Panel Data

The first component of our analysis is to construct estimates of the cross-sectional distribution child educational investment at three different points in time 1974, 1996, and 2012 using the MHSS in 1996 and 2012 and the 1974 census data for

Figure 1: Sampling Probabilities with Household Recombination



the set of houseohlds that are antecedents to the MHSS 1996 households.[1] An antecedent to a 1996 household is defined as any household that contains a member of MHSS 1996 household (zero order link), contains a member who is, between 1974 and 1996, coresident with a member of MHSS 1996 (first order link), or contains a member who is coresident with a first order link before that first order first (over the 1974 1996 interval) lives with the MHSS 1996 member.[2]

The general problem is illustrated in Figure 1. The top row of the figure denotes a set of period t households divided into two baris. The bottom row shows the households at time t+1 with the arrows connecting the two rows showing the movement of individuals over time from period t to period t+1 households. Suppose the sampling strategy is to sample one of the two strata with probability 1/2 each and then within the strata to select one household. Then the probability of sampling, for example, the household in the upper left corner is 1/6. These probabilities in combination with the household descent mapping yields a probability that each of the period t+1 households will appear in the sample. Both

[1]While we have full access to the 1974 census, we tie our hands in this way because in some cases one may have to rely on retrospectively collected antecedent data. We do compare the weighted estimates with the population data to assess the efficacy of the weights we develop.

[2]Note that the prior links described here are based on coresidence only rather than coresidence and kinship while the 2012 sample links were defined based on both coresidence and kinship. Budget constraints made it necessary to limit the 2012 in this way. Exclusive use of kinship links creates a selectivity problem, particularly in the early years of the DSS, because kinship in the MHSS database is generally only known if two individuals were coresident during at least one census.

sets of probabilities are noted on the diagram. Now imagine that one's sample of t+1 households consists of all the descendants from the one household picked by the sampling scheme. Then the weights needed to create an unbiased estimate of some t+1 characteristic are dependent not only on the sampling probability of the particular household picked at period t but of any household that could have been picked in period t with a link to the t+1 household. The same is true in reverse if the t+1 sample is representative and one constructs a period t sample from the set of antecedents of the t+1 sample, which is the case of the data we are working with.

It is evident from Figure 1 that to construct appropriate weights one needs two things. The first is a set of sampling probabilities for the period t households. The second is a set of mappings (the lines in the diagram) that indicate which period t+k households, for some interval k, would be selected in the follow up sample if a given period t household were sampled. The construction of the latter requires, in turn, two things. First, a partition H_s of individuals into households at each point s in the interval t to t+k and second, a mapping P_s that links an individual at each point of time to that same individual at a subsequent point in time.

Thus, for example, assume you have four people (a, b, c, d) distributed across two households (α, β) at time one and five (partially overlapping people) (b, c, d, e, f) into three households $(\delta, \gamma, \epsilon)$ at time two. Then

$$H_1 = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cc} \alpha & \beta \end{array} \\ \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \end{array} \tag{1}$$

and

$$H_2 = \begin{array}{c} \\ b \\ c \\ d \\ e \\ f \end{array} \begin{array}{c} \begin{array}{ccc} \delta & \gamma & \epsilon \end{array} \\ \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0. \end{pmatrix} \end{array} \tag{2}$$

Further

$$P_1 = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{ccccc} b & c & d & e & f \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array} \tag{3}$$

Then using linear algebra we may describe the coresident matrix C_s

$$C_s = H_s * H_s^T \tag{4}$$

as a matrix that has a 1 if the column person is coresident with the row person in period s. Further

$$LI_s = C_s * P_s * C_{s+1} \tag{5}$$

describes the links of people at time s to people at time $s+1$ based on coresidence in each period. Finally, we can string together the LI_s matrices

$$LH_1 = H_1^T * LI_1 * LI_2 * ... * LI_{t-1} * H_t \tag{6}$$

to characterize the links between households at time 1 to households at time t.

In our particular example,

$$C_1 = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \end{array} \tag{7}$$

$$LI_1 = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{ccccc} b & c & d & e & f \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \end{pmatrix} \end{array} \tag{8}$$

and

$$LH_{12} = \begin{array}{c} \\ \alpha \\ \beta \end{array} \begin{array}{c} \begin{array}{ccc} \delta & \gamma & \epsilon \end{array} \\ \begin{pmatrix} 2 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix} \end{array} \tag{9}$$

Thus α is linked to households δ and γ but not ϵ.

This structure of houseohld links can then be used to define corrective weights to measure the mean or other summary statistic at a particular point in time using a panel survey that is cross-sectionally representative and for which there is a previous round that is defined through coresidence. [3]. In order to formally define the corrective weights some additional notation is necessary. Formally, define

- I_t is the set of households i at time t
- I_{tx} is the set of households i at time t of type x
- $A : I_{t+1} \Rightarrow I_t$ where $A(K)$ is the set of households in I_t that contain antecedents of members of household $K \subset I_{t+1}$
- J_{t+1x} is the set of households j at time $t+1$ such that $A(j) \subset I_{tx}$
- N_{tx} is the number of households i at time t of type x
- S_t is the sample drawn at time t

Note that A can be extracted from the matrix LH defined above. That is $q \in A(p)$ if $LH[q, p] > 0$.

For some outcome c_{it} among a population with characteristics x, the population average is

$$\bar{c}_{tN} = \frac{1}{N_{tx}} \sum_{i \in I_{tx}} c_{it}. \tag{10}$$

This expression can then be reformulated so it only makes use of data available in the descendant sample.

$$\hat{c}_{tN} = \frac{1}{N_{tx}} \sum_{i \in I_{tx}} c_{it} \frac{\mathbb{1}(i \in A(S_{t+1}))}{\mathbb{E}(\mathbb{1}(i \in A(S_{t+1})))}. \tag{11}$$

We can obtain greater precision by further accounting for the ratio of the estimated population to the actual population, which is

$$\hat{i}_{tN} = \frac{1}{N_{tx}} \sum_{i \in I_{tx}} \frac{\mathbb{1}(i \in A(S_{t+1}))}{\mathbb{E}(\mathbb{1}(i \in A(S_{t+1})))}. \tag{12}$$

[3]The same basic approach applies for a subsequently collected round as in the 2012 data used here

This ratio

$$\operatorname*{plim}_{N\to\infty} \frac{\hat{c}_{tN}}{\hat{i}_{tN}} - \bar{c}_{tN} = 0. \tag{13}$$

then converges in probability to the sample average. $\mathbb{E}(\mathbb{1}(i \in A(S_{t+1})))$ can be constructed by simulation using the sampling procedure, the frame from which the sample was drawn, and all antecedent links for these households.

While these analytics reflect, except for the use of siumulation, a conventional approach to constructing sampling weights, they help to illustrate that one obtains consistency even when household coresidence, and therefore the sampling weights, are endogenous with respect to the outcome of interest. In essence the difference between the population average and the weighted sample average under recombination is only dependent on the process of sampling, over which the researcher has full control, not the patterns of coresidence. Note this importantly distinguishes weights based on recombination from weights designed to address migration or attrition.

In short, in the Matlab context, in order for the sample of antecedent households to be representative of the population, it would have been necessary to randomly select baris from the 1974 population. Instead, by randomly selecting descendants from the 1996 population for the 1996 sample without taking into account how many of them came from each 1974 household, the MHSS team inadvertently exposed the sample to the outlined potential sources of bias. [4] By sampling baris rather than households, the bias could be mitigated because baris tend to be made up of households that are linked by kinship. Yet, women are likely to join the bari of their husband upon marriage, so a 1974 household with several daughters would have descendants in several baris. In addition, the decision of some descendants to split and form their own bari or to join a different bari could also be dependent on the observable and unobservable characteristics of the 1974 household. This dynamic would again affect the probability of a 1974 household being represented in the sample.

Of course, the extent to which this is a problem depends on the the density of

[4] One of the coauthors was on the original MHSS team and now recognizes the issues with the way the sampling was done, but at the time, the focus was on getting a representation of kin networks, which were assumed to be manifested in the bari structure, without considering the endogeneity of how kin networks might spread to other baris due to the formation and recombination of households.

Table 1: Number of Links Between Selected 1974 and 1996 Households

	1996 Households										
	A11*	B11	C11*	D11	D12	D13	E11	C12	F11	B12	B13
1974 Households											
A01	3	4	0	0	0	0	0	0	0	0	0
D01	3	4	7	5	4	4	7	4	5	1	4
G02	3	4	0	0	0	0	0	0	0	0	0
D02	0	0	7	5	4	0	0	4	5	3	4
D03	0	0	1	3	3	0	0	0	5	0	0
C01	0	0	7	0	0	0	0	4	0	0	0
C02	0	0	7	0	0	0	0	4	0	0	0
E01	0	0	0	0	0	1	7	0	0	0	0
H01	0	0	0	0	0	0	0	0	0	1	2
J01	0	0	0	0	0	0	0	0	0	3	0
K01	0	0	0	0	0	0	0	0	0	3	4
G01	0	0	0	0	0	0	0	0	0	3	4
L01	0	0	0	0	0	0	0	0	0	3	0
B01	0	0	0	0	0	0	0	0	0	1	2

*Household was in the 1996 MHSS

the matrix LH, or its binary equivalent. If it is a simple identity matrix or if each column (descendant household) has only a single positive row (antecedent household) then there will be no bias at all. At least in the present context the matrix turns out to be quite dense. To illustrate this point we randomly selected one 1996 household and found its antecedents. We then took the antecedent households and constructed each of their descendants. We then repeated this process. The resulting matrix of households is presented in Table 1. Each 1974 household has multiple descendant households and each 1996 household has multiple antecedent households. Indeed looking across the 23,913 households in the 1974 census that are linked to at least one 1993 household, we find that households have on average 3.62 (sd 3.04) descendant households living in the Matlab HDSS area but only .254 (sd. .529) descendant households that were part of the sample. Conversely the 34,365 households in the 1996 census have on average 2.35 (s.d. 1.66) antecendents. It seems likely given the variation in the number of descendant households that the actual sampling probabilities for different 1974 households will be quite different from the sampling probabilities based on the 1996 sample.

These antecedent-descendant links can be used to calculate the probability that a particular 1974 antecedent household has a descendant that appears in the 1996 sample. In the case of a simple random sample this calculation is straightforward.

The probability that any particular 1974 household is picked is just $(1-(1-p)^n)$ if a 1974 household has n descendant households and the probability that any particular descendant household is picked is p. In the present case, however, because of the bari level sampling, the probability that particular households are picked is negatively correlated within baris with the extent of this correlation depending on bari size. Due to this complication in sampling, we calculate the probabilities by replicating the procedure implemented in 1996 to pick the sample based on the 1993 census. Note that this procedure would work for any arbitrarily complex sampling procedure and descent definition.

To implement the pocedure described in eqautions 11 and 12 we randomly picked 2,883 baris from the total 7,440 baris in the 1993 census, and then picked one household at random from each bari. This creates a sample of 2,883 households that is an alternate MHSS 1996 sample. We did this 100,000 times. The antecedent-descendant links were used to establish which 1974 households were represented by at least one descendant household in each sample. The probability of a 1974 household being represented in an arbitrary 1996 sample $\mathbb{E}(\mathbb{1}(i \in A(S_{t+1})))$ is the number of samples in which the household has at least one descendant out of 100,000 possible samples. Following the procedure outlined above, we created probability weights by taking the inverse of the calculated probability and assigning that as the weight to each 1974 household.[5] Having assigned each 1974 household a weight, we created a sample of 1974 households that is linked to the actual 1996 sample by taking all of the antecedents of the actual MHSS 1996 sample. This consists of 5,319 households that all have at least one descendant in the actual 1996 sample. Using the 1974 probability weights calculated earlier, we then can get a representation of the full 1974 population.

The top panel of Table 2 shows the mean value for a number of variables in 1974 for the full population as well as for the sample both weighted and unweighted. In the bottom panel we present p-values for a comparison between the population means and the differently weighted samples. For all the variables, the weighted sample is representative of the full population. The unweighted sample, on the other hand, has significantly different means for every variable. This implies that

[5]There were 3,825 households in 1974 that did not have any descendants in 1993. In this current paper we only focus on the 1974 households which have a descendant in 1993 because we cannot follow up those 3,825 households, although it is possible to examine and compare their characteristics with those of the households that do have descendants in order to determine whether their omission causes a bias.

Table 2: Different 1974 Sample Weights Compared to the Full 1974 Population

	(1)	(2)	(3)	(4)
	Mean Values for 1974 Population and Weighted Samples			
	Full 1974 Population	Our Resampling Weights	No Weights	Propensity Score Weights
Highest Edu	4.136	4.215	4.254	4.180
Number of Cows	1.158	1.179	1.355	1.173
Edu of Head	2.272	2.270	2.130	2.285
Age of Head	45.73	45.63	46.67	45.85
Household Size	6.071	6.090	6.777	6.164
Num of Rooms	1.219	1.210	1.281	1.214
Observations	24,788	5,319	5,319	5,309
Weights	24,788	24,029	5,319	24,594
	P Values for Difference between Full Population and Sample			
		Our Resampling Weights	No Weights	Propensity Score Weights
Highest Edu		0.349	0.036	0.432
Number of Cows		0.528	0.000	0.504
Edu of Head		0.978	0.002	0.787
Age of Head		0.784	0.000	0.549
Household Size		0.718	0.000	0.012
Num of Rooms		0.407	0.000	0.477

as expected, the 1996 sample is not linked to a representative set of 1974 households, and instead certain types of households were more likely to be represented in the 1996 sample. The unweighted sample has a higher average family size, which seems intuitive because a household with more family members is likely to have more descendants. On average, the households also have more cows and more rooms, both indicative of higher wealth. It seems that the 1996 sample is linked to a distinct set of 1974 households, that among other things are wealthier on average than the population, but using the weights calculated, we are able to make the sample representative of the 1974 population.

The weights that we have devised require knowledge of the full set of linkages between 1974 and 1996, which in most analogous circumstances would be unknown. In that case, a different technique that could be used, and that we compare to our weights, is to create propensity score weights. For these weights we only require a full census with some basic information and the identity within that census of the 1974 antecedents.

In particular we calculate propensity scores for the probability of inclusion in the 1974 sample and compare them to the weights assigned using the sampling

Table 3: Propensity Score Regression

VARIABLES	Dep Var: Dummy=1 if Linked to MHSS1
Highest Edu in Household	-0.0106
	(0.00654)
Household Size	0.130***
	(0.00691)
Articles Owned	-0.000998
	(0.00180)
Number of Cows	0.0165
	(0.0110)
Number of Boats	0.00637
	(0.0289)
Edu of Head of Household	-0.0256***
	(0.00704)
Age of Head of Household	-0.00124
	(0.00126)
Constant	-1.949***
	(0.0615)
Observations	24,757

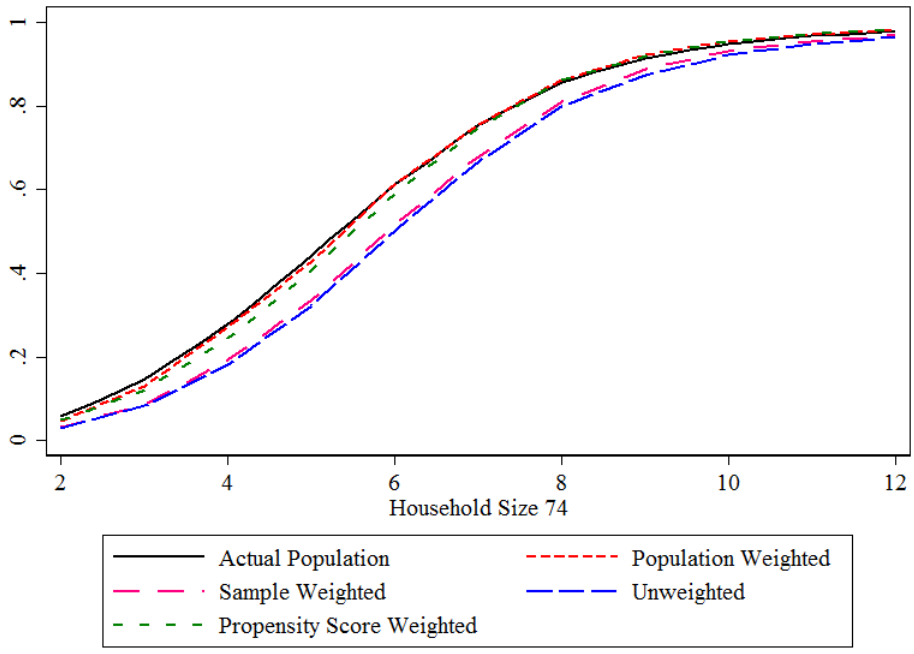
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

method. In order to do this, we take our sample of 1974 households linked to descendants in the 1996 sample and assign them a value of one for being linked, while all other 1974 households get a value of zero. This variable is the dependent variable in a logit regression. The controls used are observable characteristics of the 1974 households including highest education of anyone in the household, number of cows, number of boats, education of the head of household, the age of the head of household, the size of the family and the number of items owned by the family.[6] Table 3 shows the results of the regression used to calculate the propensity score. The coefficients from the regression are used to calculate predicted values for each 1974 household, which are equivalent to the probability that a certain 1974 household is likely to be linked to the 1996 sample based on their observable characteristics. The weight is the inverse of this predicted probability.

The last column of Table 2 shows the results of the propensity score weighting,

[6]Information was collected on ownership of a lep, harrican, watch, radio and receipt of remittance. We also conducted the propensity score analysis with all of these variables as well as including the number of descendants. In a regular propensity score analysis this variable would not be available, but given that we have a full census in 1993, we have it and tried using it to see if the accuracy of the propensity score weights increased. There is no significant difference between the weights using the number of descendants variable and those not using it, so we only show the results for the propensity score weighting procedure where we do not include number of descendants.

Figure 2: Estimated Distribution of 74 Household Size with Various Weights



which can be compared to the sample weights, the scenario with no weights and the actual population means. Both our sampling weights and the propensity weights come very close to approximating the true population means. For most of the variables, the propensity weights are not significantly different from the population means, but they do differ significantly in the case of family size. Although the propensity weights are fairly representative of the full population, the difference in family size is worrisome because it could mean there are other unobservables that are also significantly different from the population averages. Therefore, our weights yield the most representative sample weighting structure.

A visual perspective on the effects of different weighting schemes is provided by Figure 2, which compares the cumulative distributions for the 1974 family size for the different weighting schemes. In addition to the basic measures presented in Table 2 (actual population, population weighted, unweighted and propensity score weighted) we add a measure of the estimate one would obtain if one used only the sampling weights from the actual MHSS 1996 sample (as opposed to the set of all possible 1996 samples) to reweight the 1974 sample. This turns

out to be a straightfoward calculation because one household per bari is selected and sampling is independent across baris. In particular, disregarding all non-sampled households, the probability a particular antecedent household is picked is $(1 - \prod_j (1 - p_j))$ where j indexes all the 1996 households in the MHSS sample that are descendants of a particular 1974 household.[7]

The results are quite striking. As suggested by Table 2 the distribution for the actual population is very closely approximated by the distribution based on the weights simulated by redrawing the sample. The propensity score does not do as well but is again quite close. On the other hand the sample weighted estimates coincide very closely with what one gets based on the unweighted data and both are quite far from the actual population. This result can be attributed to the facts that one household per bari was picked in the 1996 sample and that many of the descendants of a particular 1974 household tend to live in the same bari. As a result the probabilities calculated from any particular sample gives very little sense of the likelihood of a particular 1974 household being picked across multiple draws of the sample. Specifically consider a household with five descendant households in one five household bari versus a household with 1 descendant household in a five household bari. If that bari is sampled 100% of the time then in the former case the 1974 household will be picked 100% of the time but in the latter case it will only be picked 20% of the time. But since any given 1996 household is picked 20% of the time the sampling weight will assign the same weight to the 1974 houseohld regardless of which of these two scenarios obtains.

To complete an analysis of changing inequality we need comparable measures for 1996 and 2012. Calculating the distribution of an outcome in 1996 is, of course, straightforward because we have a stratified representative sample in 1996. To calculate the distribution in 2012, an obvious approach would be to start with the 1996 sample and carry the weights forward in the descendant households. This approach is potentially problematic for two reasons. First, it is not so clear what one should do if a 2012 household has multiple antecedents. This is of course very unlikely in a small sample of a large national population. However, given the relatively high sampling probability here (10%) and the number of people

[7]Note that if each 1974 household were linked to only one household in the sample, then the probability assigned in this way would just be the probability that the linked 1996 household is in the sample. These weights do not account for the fact that a 1974 household could theoretically enter the sample through any of its descendants, but only look at the particular descendants that it did enter the sample through.

Table 4: Estimates of Family Size in 2012 by Weighting Scheme

	Full Population	Random Smpl All Desc.	Random Smpl One Desc per 1974 HH	MHSS1 Linked Smpl 1996 Wts	MHSS1 Linked Smpl 2012 Wts
Avg Family Size 2012	4.394795	4.443986	4.542926	4.575445	4.398142
Obs	49,988	25,783	4971	4671	4681

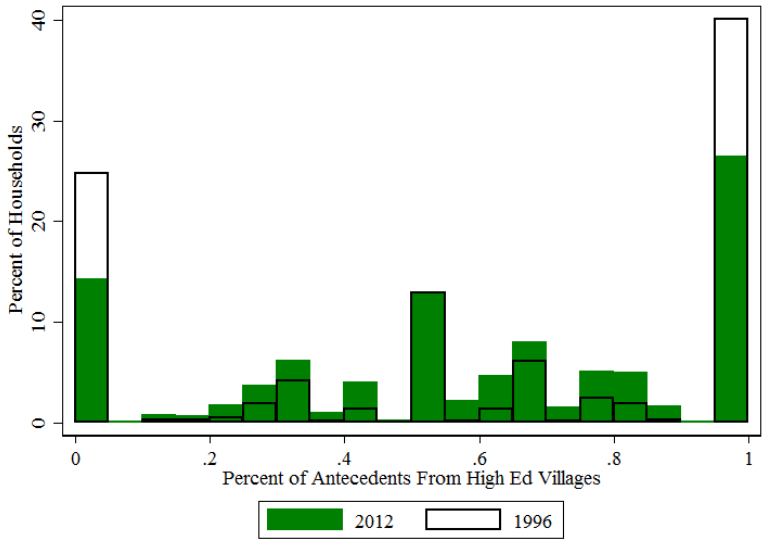
in each household it is not at all unlikely here. If the antecedent households had been sampled independently with probably p then an appropriate adjustment might be to use a weight of $1/(1-(1-p)^2)$, but, of course, sampling was not independent. Second, as has been emphasized throughout this would only account for antecedents in the 1996 MHSS sample. There are also likely to be antecedents not in the sample that affect overall sampling probabilities. Again, however, we can use simulation to redraw the 1996 population coupled with the antecedent/descendant matrix from the DSS data to reweight appropriately, this time linking the sample to the 2012 descendants. Family size by various types of weighting schemes appear in Table 4. While the differences are smaller than for 1974, we again see that the formally derived weights are the closest match to the full population.

4 Changes in educational inequality

We now turn to the question of changes in educational inequality over the 1974-2012 period. For this purpose we use the population level data from 1974, as well as the 1996 and 2012 survey data, appropriately weighted, to characterize educational inequality, overall economic and demographic mobility and to provide a benchmark that will serve to evaluate the effectiveness of the different weighting schemes.

We start by constructing an indicator that permits us to aggregate educational attainment across ages within households and to compare changes over time. Making use of the education data we determine the mean and standard deviation at each age of completed schooling for children aged 6-16 in the 1996 MHSS using the household cross-sectional weights. We then subtract from each individuals education, in the 1974, 1996, and 2012 data, the mean schooling and divide by

Figure 3: Distribution of Fraction of Antecedent Households from 1974 High Ed Villages Among All Households in 1996 and 2012

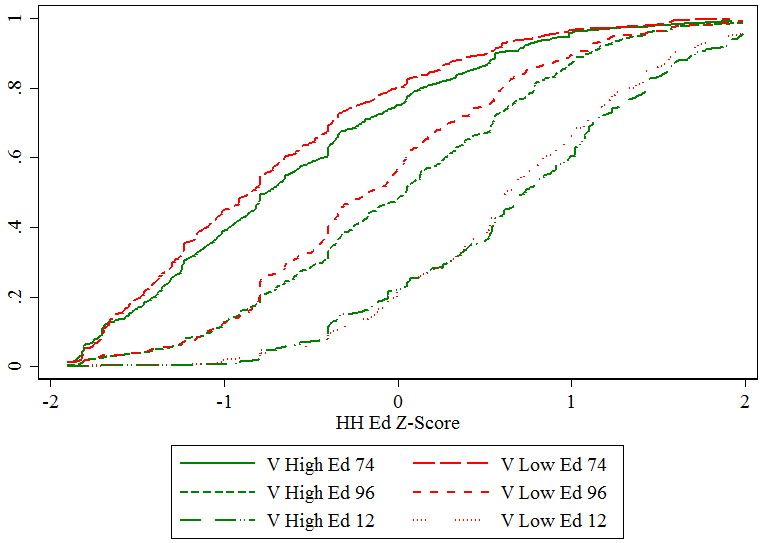


the standard deviation of schooling from 1996 to create an individual educational z-score. We then aggregate at the household level and year to obtain, in effect, a household level z-score of child education that reflects the relative intensity of child education of each household.

To capture the persistence of spatial inequality across the study area we also created a measure of initial village stratification based on village literacy. In particular, a 1978 survey contains information on the fraction of each village in which the head is literate. A high education village is one in which at least 40% of the heads in the village were literate at that time.

A complication that arises in stratification by village, and indeed a possible source of convergence over time, is the extent to which households in 1996 and 2012 are composed of individuals from high and low literacy villages as determined in 1974. In particular, Figure 3 plots the distribution, for each household in 1996 and 2012, of the fraction of 1974 antecedent households in the study area in a high literacy village. By 1996, 66% of households had only one type of antecedent with the remainder being of mixed "ancestry". By 2012, however, this fraction had

Figure 4: Distribution of HH Ed Z-score by Year and Village Ed



dropped to 40.

Plots of the cumulative distribution of the educational index by village literacy in the three years are presented in Figure 4. The most striking feature is the secular increase in child education over the period. Using the 1996 distribution as a base, in the high literacy villages there is a 0.6 standard deviation increase in education between 1974 and 1982 and a somewhat larger 0.7 standard deviation increase between 1996 and 2012. There was little reduction in inequality however at least as measured by the interquartile range. In particular in 1974 the interquartile range was about 1 standard deviation but had increased by 1.2 standard deviations by 2012. Not surprisingly we also see that there is a small gap in educational investment between more and less literate villages in 1974. Interestingly, this gap widens somewhat by 1996 but by 2012 is almost gone. This pattern will be even more clear when we look directly at mobility in educational investment.

5 Correcting for Bias in Descendant Selection

The second problem with using the data without corrective weights is also related to the bari structure of the sample. Because baris were the unit that was randomly sampled, and only one household was picked from each bari, if two descendants from the same 1974 household were in the same bari, they would never both be picked to be in the sample. If, on the other hand, two descendants from the same 1974 household were in different baris, then it is possible that both could be picked for the sample, and even more so if they are in small or single household baris. We already mentioned how this could affect the representativeness of the 1974 population if the characteristics of 1974 households are correlated with the decision of their descendants to stay in the same bari or split off. In addition, if the decision to stay in the same bari as other descendants, split off into a new bari or join a different bari is correlated with the characteristics of the 1996 households, then it is not possible to accurately estimate average descendant outcomes, distorting the results of intergenerational analyses.

To illustrate the problem, suppose a 1974 household has three descendants and we are interested in the average outcome of descendants from this household. We need an accurate estimate of the average outcome for the descendants of the household, but in most cases all three descendants will not be in the sample. Now suppose that certain attributes determine whether households remain in the same bari or split off. For example, it is possible that the poorest descendant household of a family might choose to split off and look for better opportunities in a different location, while the two richer households remain in the same bari because they are already well off and would not want to leave their land, assets, network, etc. If this behavior were systematic in the population, it would mean that two high-outcome descendants, for example, would never both be in the sample, instead there would tend to be a richer and a poorer descendant. Thus, if the average outcome of descendants for particular households is calculated by taking the arithmetic average of the descendants that show up in the sample, then the sample will consistently underestimate the true average outcome of descendants.

If household recombination is random, so that the probability of getting any combination of households with certain attributes is equal then there should be no bias, and the arithmetic average will be the average outcome for the descendants. This seems unlikely though, given that certain attributes such as wealth have been shown to play some role in household recombination.

To understand the problem more clearly, imagine the following scenario. Suppose there are two descendants from a 1974 household with 1996 outcomes c_1 and c_2. Theoretically we could see just c_1, just c_2 or both in the sample. In this example, household c_1 is never picked alone and there is a .5 chance of picking both c_1 and c_2 and a .5 chance of picking just c_2.[8] If we were to take the average of the households if they show up together and take the value of c_2 when we only have c_2, then we get the following expected outcome:

$$\begin{aligned} \mathbb{E}(c) &= 0 * c_1 + \frac{1}{2}\left(\frac{c_1 + c_2}{2}\right) + \frac{1}{2}c_2 \\ \mathbb{E}(c) &= \frac{1}{4}c_1 + \frac{3}{4}c_2 \neq \frac{1}{2}c_1 + \frac{1}{2}c_2 = \bar{c} \end{aligned} \tag{14}$$

In expectation, we are not getting the average for the descendants. If the probabilities were random with respect to the outcome of the households, then with a large enough sample, this inconsistency would average out. The problem arises if the probability is correlated with attributes of the 1996 households. For example, if the data consisted of two descendants for every antecedent and the probability of being picked is correlated with an outcome of interest so that c_2 is always the lower-outcome antecedent and c_1 is always the higher-outcome one, this would result in a lower estimate of the average outcome for descendants.

There are various ways to tackle this problem. In this case we could take c_2 if we only have c_2 and only take c_1 if we pick both c_1 and c_2, which would give us an expectation equal to the average. Yet in doing that, information on descendant c_2 would be thrown out if both descendants are in the sample. In addition, this is a solution for this particular set of probabilities. There are also more complicated probabilities in the data where there is a probability of seeing all three combinations.

Formally, we propose a method based on extension of the method used above to develop the 1974 measures. We then show that this method can be derived

[8]This scenario might seem unlikely, but its purpose is to illustrate the more complicated case of several descendants where some live in the same bari and therefore the probability of picking two descendants living in the same bari is 0. Doing the example with more descendants makes the calculations messier and detracts from the point of simply illustrating why it is important to consider how descendants are weighted.

based on constrained minimization that is robust with respect to variation in the relationship between sampling probabilities and outcomes. We also explore the properties with respect to a simpler but less robust procedure. In particular, we wish to estimate

$$\Delta\bar{c}_{tN} = \frac{1}{N_{tx}} \sum_{i \in I_{tx}} \frac{1}{|A^{-1}(i)|} \sum_{j \in A^{-1}(i)} (c_{jt+1} - c_{it})^9 \tag{15}$$

As for Equation 15, we can construct an estimate of this quantity using only the sampled data and appropriate weights:

$$\Delta\hat{c}_{tN} = \frac{1}{N_{tx}} \sum_{i \in I_{tx}} \frac{1}{|A^{-1}(i)|} \sum_{j \in A^{-1}(i)} (c_{jt+1} - c_{it}) \frac{\mathbb{1}(j \in S_{t+1})}{\mathbb{E}(\mathbb{1}(j \in S_{t+1}))} \tag{16}$$

It again follows that

$$\operatorname*{plim}_{N \to \infty} \frac{\Delta\hat{c}_{tN}}{\hat{i}_{tN}} - \Delta\bar{c}_{tN} = 0. \tag{17}$$

Note that (16) does not depend on the 1974 sampling probability. However, by dividing by this probability just after the first sum and multiplying by this probability after the second sum it can be seen that the appropriate estimates of average growth by initial characteristics is obtained by summing across descendant households in the sample that come from each 1974 household the scaled change in the outcome

$$(c_{jt+1} - c_{it}) \frac{\mathbb{E}(\mathbb{1}(i \in A(j)))}{|A^{-1}(i)|\, \mathbb{E}(\mathbb{1}(j \in S_{t+1}))} \tag{18}$$

and then combining the different 1974 households using weights derived from the resampled 1974 probabilities.

We now show how this estimate can be derived from a constrained minimization. The first criteria we want to meet is that any weights should lead to an outcome where the expected value for descendants' outcome is equal to the actual mean of the outcomes. The way this would look in the case of two descendants with outcomes c_1 and c_2 and conditional on observing at least one 1996 household with the probability p_a of picking just household 1 in the sample, probability p_b of picking both household 1 and household 2, and probability p_c of picking only household 2 in the sample is as follows:

[9] $|A^{-1}(i)|$ denotes the size of the set of households descending from i.

$$\mathbb{E}(c) = p_a w_a c_1 + p_b(w_{b1} c_1 + w_{b2} c_2) + p_c w_c c_2 = \frac{1}{2} c_1 + \frac{1}{2} c_2 = \bar{c} \tag{19}$$

where w_a, w_{b1}, w_{b2}, and w_c are the weights applied to outcome 1 if it is only observed, outcomes 1 and 2 if both are observed, and outcome 2 if only 2 is observed, respectively.

There are many different possible ways of weighting the observations in order to get an expected value for descendants equal to the average. Some of these weights might lead to a large variance in the estimates, depending on the probabilities. Therefore, in addition to getting the correct mean income in expectation using our weights, we also want to minimize the variance from different probabilities of descendant combinations being picked for different antecedents. We want to minimize the following:

$$\begin{aligned} Z = [var(p_a)(w_a^2 c_1^2) + var(p_b)(w_{b1} c_1 + w_{b2} c_2)^2 \\ + var(p_c)(w_c^2 c_2^2) - 2cov(p_a, p_b)(w_a c_1)(w_{b1} c_1 + w_{b2} c_2) \\ - 2cov(p_a, p_c)(w_a c_1)(w_c c_2) - 2cov(p_b, p_c)(w_{b1} c_1 + w_{b2} c_2)(w_c c_2)] \end{aligned} \tag{20}$$

Both equation 19 and equation 20 depend on the actual mean and variance of the outcome values, but we do not have all of the outcome values. Therefore, the weights must work more generally and not be sensitive to the mean and variance of the outcomes. Again, this could be achieved in different ways. A sufficient condition for equation 19 to hold is that it holds for all outcomes, which can be ensured by taking derivatives with respect to the income values c_1 and c_2. In our simple two descendant example, this yields the following two equations, both of which need to hold in order for our first condition to be met:

$$\begin{aligned} p_a w_a + p_b w_{b1} &= \frac{1}{2} \\ p_b w_{b2} + p_c w_c &= \frac{1}{2} \end{aligned} \tag{21}$$

We apply a similar logic to our second condition and take second derivatives with respect to c_1 and c_2 in order to come up with the following objective function:

$$\min_{w_a, w_{b1}, w_{b2}, w_c} \frac{d^2 Z}{dc_1^2} + \frac{d^2 Z}{dc_2^2} \tag{22}$$

This ensures that the variation in the fraction of households in each sample has a small impact on the computed average income for descendants because the weights are applicable and minimize variance no matter what the actual incomes are. We calculate the weights by minimizing equation 22 conditional on equations 21. There are other, more complicated, criterion functions that could have been used, but we believe that this simple one still allows us to find weights that help to mitigate the potential bias arising from the bari structure. We will show how our weights using this procedure compare to not using weights and that indeed they can help to address the potential bias.

Solving the minimization problem, we find that the weights that minimize the variance of the estimates and in expectation yield the true average outcome are based on the probability of sampling a 1996 household. The combination in which a household appears (whether a descendant appears alone in the sample or if there are several other descendants in the sample from the same antecedent) does not affect the weight. This is surprising because as shown in our example in equation 14, the combination of descendants in a bari affects the expected value we get. Yet in trying to minimize the variance in a manner general enough to apply to all income values, the combination in which the descendants appear is no longer important. Nevertheless, the weight not only depends on the probability of being picked in 1996, but also on the total number of descendants. With the number of descendants in the denominator of the weight, those households who come from an antecedent with many descendants receive a smaller weight. Finally, the probability of the 1974 household i being represented in 1996 also factors in to account for the possibility that there is no descendant in the sample at all. Therefore, for our two descendant example the weights are:

$$w_a = w_{b1} = \frac{\Pr(i)}{2 * \Pr(1)} = w_1$$
$$w_c = w_{b2} = \frac{\Pr(i)}{2 * \Pr(2)} = w_2$$

where Pr(i) is the probability of the antecedent i of the household being represented by a descendant in the 1996 sample. We can generalize this result to assign a weight to every descendant j of a 1974 household i with M descendants:

$$w_j = \frac{\Pr(i)}{M * \Pr(j)}$$

This expression is exactly the formula derived above (Equation 16).

Without the outcomes of descendants to help assign the weights directly based on our original two specifications, it was necessary for us to come up with weights that are generalizable no matter what the outcomes might be. Given that there are a number of ways we could have devised the weights, it is important to show that using the weights improves estimates. To do this, we have conducted a simulation to demonstrate how the weights compare to not using weights.

The simulation is a simplified case of our data in order to focus on the effect of the weights when the probability of being picked is correlated with the outcome, and how the performance of the weights depends on the extent of the correlation. We do not incorporate the sampling structure but instead look at how our weights perform in the case where we have 1000 antecedents and each has at least one descendant chosen for the sample. Therefore, here our Pr(i) is equal to 1 because each antecedent has probability 1 of having a descendant in the sample.

In the simulated data, each antecedent has exactly two descendants and each of their descendants has randomly been assigned a log outcome from a normal distribution with mean 8.58 and variance 1.15.[10] This mean and variance were chosen as they were the mean and variance of one outcome in the data, log consumption for the 1996 sample.

We assigned probabilities for the following three events: household 1 is selected, household 2 is selected, both household 1 and household 2 are selected. The probabilities of being chosen are based on the random outcomes using a logistic function in order to ensure a correlation between outcomes and probabilities.

[10] Although we only do the simulation with two descendants per antecedent household, the results are generalizable to more descendants and we have done some simulations including more than two descendants, but do not include the results here. We have also done simulations where we change the variance of the income variable, and this has also not affected the general result, so we omit those results.

We varied the size of that correlation by multiplying outcomes in the logistic expression times a coefficient δ, which is manipulated. The same δ is used for all three probabilities. The three probabilities are:

$$\begin{aligned} Pr(1) &= \frac{e^{\delta * c_1}}{e^{\delta * c_1} + e^{\delta * c_2} + e^{\delta * \bar{c}}} \\ Pr(2) &= \frac{e^{\delta * c_2}}{e^{\delta * c_1} + e^{\delta * c_2} + e^{\delta * \bar{c}}} \\ Pr(1\&2) &= \frac{e^{\delta * \bar{c}}}{e^{\delta * c_1} + e^{\delta * c_2} + e^{\delta * \bar{c}}} \end{aligned} \tag{23}$$

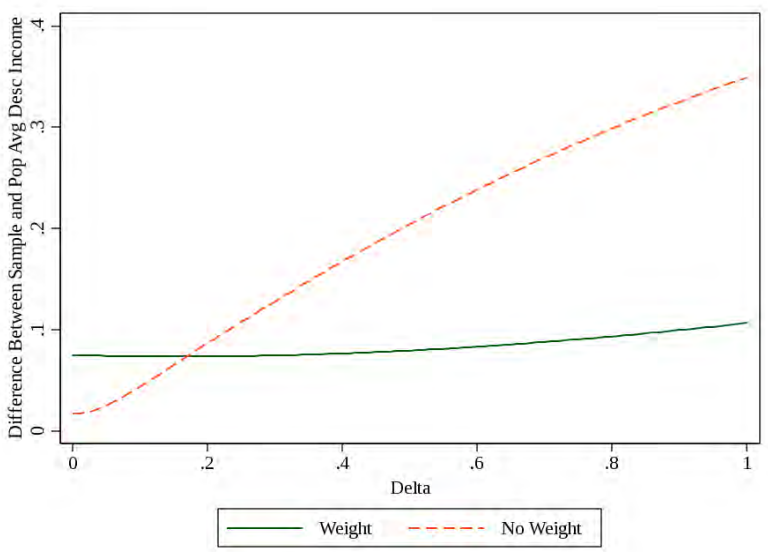
where $\bar{c}$ is the arithmetic mean of the two outcomes. The coefficient delta varies from 0 to 1 in .01 intervals. A coefficient of 0 implies that each event has a one third probability of occurring irrespective of outcome, so there is zero correlation between outcomes and probability. As the coefficient grows, the dependence between outcome and probability increases up to when the coefficient becomes 1, which gives the highest dependence between outcome and probability.

When the correlation is positive, if one household has a higher outcome than another, it will always have a higher probability of being selected alone, the probability of selecting both households will be next highest, and the probability of selecting the low-outcome household alone will be smallest. As the coefficient grows, this ordering of probabilities does not change, but the differences in probabilities become starker.

A sample of descendants is chosen based on the probabilities. One of the events is randomly chosen based on the probability of each event occurring. Depending on the combination of descendants chosen for each antecedent, the weighted mean income is calculated based on the weights ($w_j = \frac{1}{2 * \Pr(j)}$). The mean outcome is also calculated with no weights, which entails taking the arithmetic mean if both descendants are in the sample, and taking the plain value of the descendant chosen if only one is in the sample. These two means are compared to the actual mean outcome for each descendant. Actual mean outcome is subtracted from the simulated mean outcome with and without weights and the absolute value is averaged to find the mean difference between actual and sample descendant outcome for the 1000 antecedents.

In order to make sure the simulation is robust to outliers in the events picked, a set of events was chosen 500 times. The average of the absolute mean difference

Figure 5: Average absolute difference between sample and actual descendant income means for different levels of correlation



and squared error was calculated for each sample. This analysis was done with 1000 samples having different income values (and thus probabilities). This procedure was done for each δ from 0 to 1 in .01 intervals.

Figure 5 shows the sample average difference between the outcome calculated with our weights and the actual outcome, as well as the sample average difference between the outcome calculated without weights and the actual outcome.This is graphed for various deltas which represent how dependent the probabilities are on outcome. A delta of 0 signifies that the probability is not dependent on outcome at all, and a delta of 1 signifies a high degree of dependence between outcomes and the probabilities. The figure demonstrates higher variability in the average error when no weights are used versus when weights are used. Although the weights do not always perform better than not using weights, on average they are consistent in the level of error no matter what the correlation between outcome and probability, and this level is relatively low. The level of error ranges from almost none to almost .35 when no weights are used, while the error with the weights remains consistently under or close to .1.

If there is no correlation between outcome and the probability of certain combinations of households being picked, it would be better to use no weights. In that scenario, the mean descendant outcomes calculated using no weights are very close to the actual mean outcomes for the sample. When delta goes above .17, then the weights become better to use. This switch occurs for a relatively small delta, so if there is reason to believe a link exists between outcome and the probability of being chosen in the sample, then it is better to use the weights as compared to no weights.

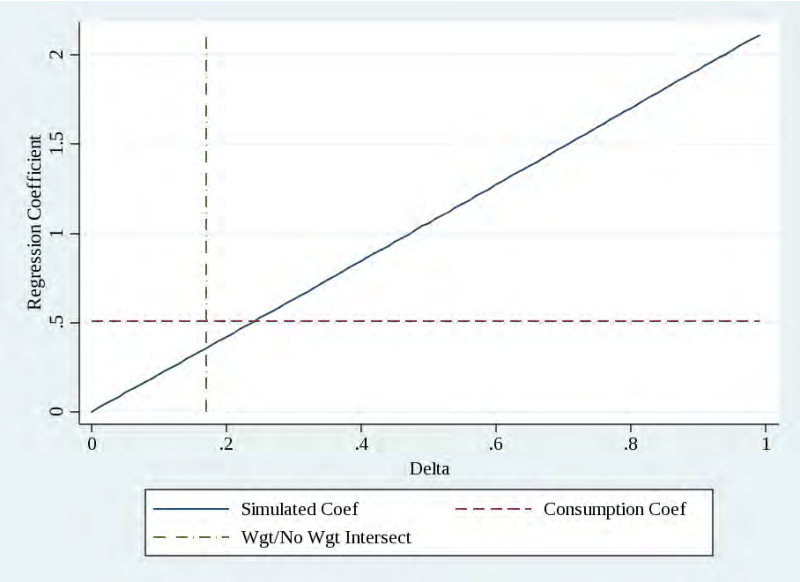
The extent of the correlation can be tested using the census conducted in Matlab in 1996 that collected data from most of the households that were present in 1996. [11]. This survey has very limited data, but it does include several variables on household assets and household infrastructure which can be used to create an index as a proxy for household income. [12] The index is based on the importance of these assets and infrastructure for predicting consumption in the MHSS. Although the MHSS does not have an income variable, it provides data on consumption, which is regressed on the assets and infrastructure variables in 1996 that are analogous to the 1996 SES ones. The coefficients from this regression gives a set of weights for how important different assets are in predicting the income of a household. These weights are applied to the asset and infrastructure variables in the 1996 SES survey and create an index based on consumption. This measure of consumption, used as a proxy for income, can be used to calculate the correlation between income and the probability of being chosen for the sample. The probability values come from the simulations done for the 1974 weights. The log probability of being picked to be in the MHSS is regressed on the log income index to get an elasticity.

To put this in the context of the δ from earlier, similar regressions were run using the simulated data. Regressions of the probability of being picked on the simulated income were run for each δ between 0 and 1 at .01 intervals. Figure 6 plots how the coefficient grows linearly as δ increases. The horizontal dashed line plots the coefficient value based on the regression using the consumption index. The vertical dash and dot line marks the δ at which the no weight and the weight lines crossed in Figure 5. The coefficient from the consumption regression crosses

[11] The MHSS sample was drawn on the 1993 census. However, we also use data from the 1996 census, because it contains various asset measures and household structure. It does not, however, included education

[12] The variables used for the index are whether a household has a cow, boat, clock, or radio; whether it gets its drinking water from a tubewell; and whether the roof is made of tin.

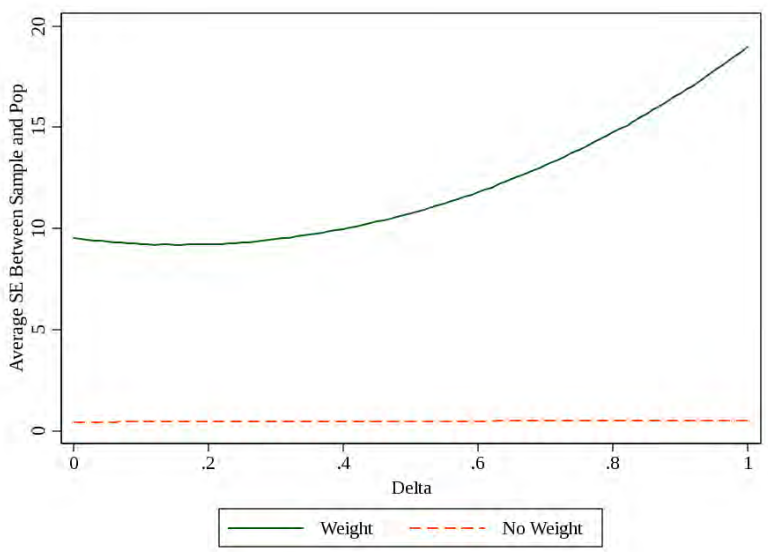
Figure 6: Coefficients of Probability Regressed on Income for the Simulated Data and the 1996 Census Data



a little above where it becomes better to use the weights. Therefore, this seems to imply that it is slightly better to use the weights in this situation, though the results with and without the weights should be relatively close.

Up to now, the discussion has focused on the performance of the weights on average for the sample. Except in the cases where there is very little correlation between income and the probability of being selected, the weights give a good approximation of the mean income of descendants on average for the sample. If we are interested in how they perform for individual households though, the average squared error is much bigger when using the weights as compared to not using weights. This is because as the sample gets bigger, the average income with the weights will be close to the expectation (in accordance with the law of large numbers), which due to the construction of the weights should be equal to the actual mean income. So with a sample of 1000, this holds true. But when looking at each individual antecedent and how the income calculated with the weights compares to the actual income, the squared difference is much bigger because the weights can cause some outliers. This is because chance means that sometimes even an event with a small probability will be picked, but that means it will have

Figure 7: Average squared difference between sample and actual descendant incomes



an extremely large and distorting weight.

Figure 7 shows how the weights compare to using no weights for individual antecedent households. The average squared error for an antecedent when not using weights is around 0.4, while the average squared error per antecedent starts out a little less than 9 and grows to almost 20 when using the weights. This is because, especially as the probabilities become more dependent on the income draws, there are more likely to be outliers with a very small or very large probability, which means a very large or small weight. Using such a large or small weight will give a more skewed average income for a particular household than if no weight is applied. Nevertheless, even though for individual antecedents there is higher variability in the mean income calculated, looking at the whole sample, the very small incomes (due to very small weights) will be balanced by the antecedents that get very large incomes due to large weights, and in this way the average income for the sample will approach the actual average income.

What this means is that it is important for a researcher to think about the type of analysis he or she is running in order to determine whether using the weights

is appropriate. In the case where one is interested in average effects, such as running regressions, using the weights would lead to more accurate results (if there is a link between the probabilities of being picked and the variable of interest). If, instead, one is interested in the effects on certain quintiles of the population, which would involve using the average income to break people up into those quintiles, then the weights would distort the data. Therefore, it is important to be aware of the goal of any analysis and how using these weights might affect it in order to make sure that the weights are used correctly and are helping to improve the accuracy of the results rather than leading to additional bias or distortions[13].

6 Household Mobility Using the Sample Weights

Before proceeding to the analysis of educational mobility, which can only be done using sampled data, we consider the application of our different weighting schemes to family size change, which has the advantage that we can compare different weighting schemes to those at the population level. We in particular categorize 1974 households by educational investment, household size, consumption growth per capita index derived from assets, and village literacy. We then regress 1974 to 1996 household size change on these categories.

The first column of Table 5 uses the population data. The second column implements our preferred weighting scheme as described in equation 16 A combination of 1974 and 1996 sampling probabilities and the total number of descendant households are used to average results across descendant households and then these are aggregated using the 1974 resampled weights. The third column is analogous to our propensity score approach and uses the predicted number of descendant households based on a regression that includes only the 1974 asset variables. The basic idea is to find an approach that would be effective in the presence of sampled data but without full information on the actual number of descendant households for each antecedent. The fourth column constructs the simple within sample average of education of the descendants but then weights using the 1974 sample weights. The fifth column constructs a weighted mean among the sampled descendants using the 1996 cross-sectional weights and then weights antecedents using the 1974 sample weights. The sixth column is the same as the fifth except

[13] Although in the empirical section we break up the 1974 data into thirds based on income, we use the full population and not the weights to do this. We only use the weights in running the regressions

Table 5: HH Family Size in 1996 by 74 Conditions and 74 Village Ed

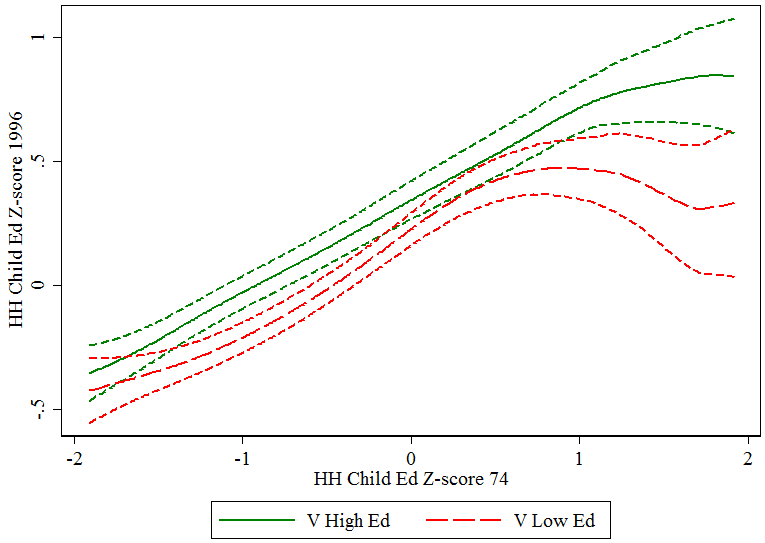
VARIABLES	(1) Population	(2) Formal	(3) Predicted	(4) 74 Weights	(5) 74/96 Weights	(6) 96 Weights	(7) No Weights
Ed Low	0.0223	0.0570	-0.118	0.127	0.138	0.0387	0.00971
	(0.0344)	(0.124)	(0.139)	(0.116)	(0.116)	(0.104)	(0.104)
Ed High	-0.110**	-0.171	0.233	-0.190	-0.191	-0.263*	-0.272*
	(0.0433)	(0.207)	(0.177)	(0.192)	(0.192)	(0.145)	(0.144)
H Size Low	2.046***	2.060***	2.118***	2.005***	2.007***	1.959***	1.956***
	(0.0335)	(0.136)	(0.128)	(0.129)	(0.129)	(0.101)	(0.0997)
H Size High	-3.411***	-3.163***	-3.565***	-3.109***	-3.139***	-3.523***	-3.489***
	(0.0526)	(0.211)	(0.270)	(0.182)	(0.182)	(0.116)	(0.116)
Cons Low	-0.597***	-0.658***	-0.610***	-0.630***	-0.629***	-0.694***	-0.695***
	(0.0391)	(0.154)	(0.160)	(0.148)	(0.148)	(0.113)	(0.111)
Cons High	0.303***	0.418**	0.496***	0.465***	0.464***	0.469***	0.463***
	(0.0394)	(0.167)	(0.158)	(0.157)	(0.157)	(0.120)	(0.117)
V High Ed	-0.181***	-0.148	-0.251*	-0.0350	-0.0344	-0.0745	-0.0725
	(0.0320)	(0.136)	(0.136)	(0.129)	(0.129)	(0.0900)	(0.0891)
Constant	-0.803***	-0.790***	-0.948***	-0.778***	-0.801***	-0.639***	-0.580***
	(0.0418)	(0.164)	(0.174)	(0.160)	(0.160)	(0.132)	(0.133)
Observations	19,820	4,690	4,688	4,690	4,690	4,690	4,690
R-squared	0.461	0.261	0.159	0.325	0.328	0.339	0.335

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

that it does not adjust using the resampled weights. Finally the seventh column works with the unweighted data. Estmates are clustered using the 1996 household to reflect the fact that each 1996 household, once sampled, can contribute multiple antecedents and thus contributes multiple observations. As we have population data, for columns (2)-(7) we report mean coefficients and standard deviations of estimates obtained by multiple draws of the sample based on the 1996 sampling scheme rather than just the coefficients and standard errors based on the MHSS 1996 sample.

As is evident from the table, the formal weighting procedure matches quite closely the results from the full population with a coefficient on village literacy, for example, of -.148, which matches well with the population effect -.181, though one cannot reject across samples an estimate of 0. The predicted estimates are about 50% larger than those based on thee population. The four other approaches yield a coefficient on village literacy that is less than half the size of the population estimates. We include as anticipated that the proposed procedure wokrs well and there is some evidence that other procedures underestimate at least inter-village differences in mobility; however, the weighting scheme also leads to a modest increase in standard deviations of coefficient estimates so inferenece is not substantially different with the formal weights.

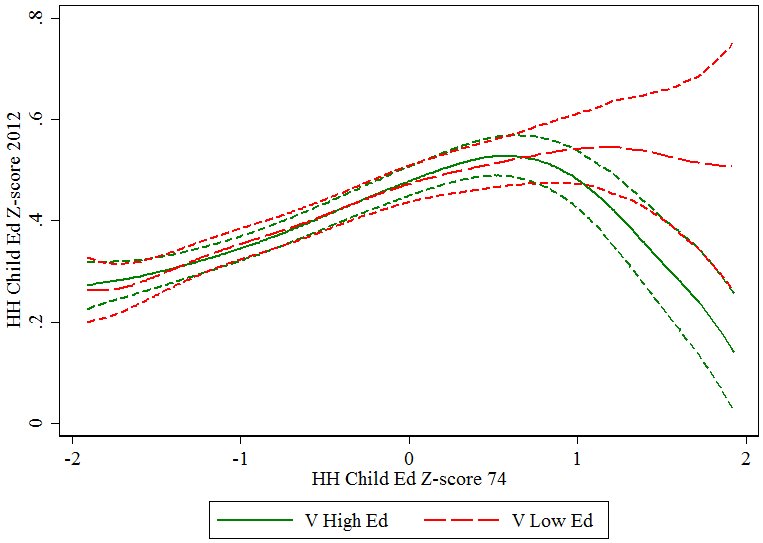
Figure 8: HH Education in 1996 by 74 HH Education and 74 Village Ed



We now turn to the analysis of educational mobility. Figure 8 presents the 1996 household average child educational z-score by 1974 education and treatment status and is constructed using the formal procedure defined above given the absence of educational attainment data in the 1996 census. The hashed lines indicate confidence intervals. We see considerable persistent in educational investment over this period as the lines slope upward. Note, however, the scale on the vertical axis. In fact a two standard deviation difference in educational investment in 1974 corresponds to a 1.2 standard deviation difference in educational investment in 1996. Thus there is evidence of a fair bit of economic mobility as well. We also see that the advantage of the higher literacy villages persists through 1996 in a fairly uniform ways.

Figure 9 is also upward sloping through most of its range. There are two noticeable differences, however. A two standard deviation in education in 1974 translates into only a .5 standard deviation in 2012 among descendant households. Moreover there are by 2012 no discernible differences between the descendants of the more and less literate villages in 1974. Overall there has been a high degree of convergence over this 38-year period at both the household and village level,

Figure 9: HH Education in 2012 by 74 HH Education and 74 Village Ed



suggesting that this has been a period of considerable educational mobility.

Tables 6 and 7 provide regression results that correspond to the previous two figures and also permit comparison between our preferred weights and the other alternatives explored above. The coefficients do not differ markedly. The village literacy coefficient, for example, is .144 in the preferred weighting scheme. The propensity score yields an estimate that is 10 percent higher, while the other weighting schemes are 10 to 30 percent lower. The 2012 estimates yield a similar conclusion with the various results yielding a pretty clear zero relationship between village literacy and educational investment in 2012. Consistent with the results from the household size specification above, the weighting schemes do not yield qualitatively different specifications but are somewhat different. Of course, whether and how these results differ will depend in general on patterns of household recombination and sampling schemes and thus it is not clear how general this result might be.

Table 6: HH Ed Z-score 96 by 74 Conditions and 74 Village Ed

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Formal	Predicted	74 Weights	74/96 Weights	96 Weights	No Weights
Ed Low	-0.226***	-0.270***	-0.244***	-0.234***	-0.244***	-0.265***
	(0.0601)	(0.0592)	(0.0558)	(0.0555)	(0.0314)	(0.0327)
Ed High	0.394***	0.298***	0.395***	0.398***	0.293***	0.288***
	(0.0831)	(0.0833)	(0.0690)	(0.0690)	(0.0386)	(0.0392)
H Size Low	-0.144**	-0.0716	-0.183***	-0.179***	-0.130***	-0.143***
	(0.0615)	(0.0582)	(0.0581)	(0.0580)	(0.0365)	(0.0370)
H Size High	0.0391	0.117*	-0.00203	-0.00748	0.0515	0.0596*
	(0.0701)	(0.0646)	(0.0615)	(0.0611)	(0.0329)	(0.0342)
Cons Low	-0.175***	-0.149***	-0.179***	-0.176***	-0.153***	-0.163***
	(0.0602)	(0.0527)	(0.0582)	(0.0579)	(0.0337)	(0.0350)
Cons High	0.192***	0.224***	0.166**	0.162**	0.170***	0.178***
	(0.0712)	(0.0663)	(0.0664)	(0.0662)	(0.0395)	(0.0408)
V High Ed	0.144***	0.157***	0.130**	0.125**	0.104***	0.114***
	(0.0546)	(0.0522)	(0.0504)	(0.0503)	(0.0300)	(0.0307)
Constant	-0.0428	-0.130*	0.0164	0.0136	-0.00424	0.00312
	(0.0700)	(0.0685)	(0.0667)	(0.0665)	(0.0435)	(0.0445)
Observations	3,469	3,467	3,469	3,469	3,469	3,469
R-squared	0.064	0.028	0.128	0.127	0.103	0.108

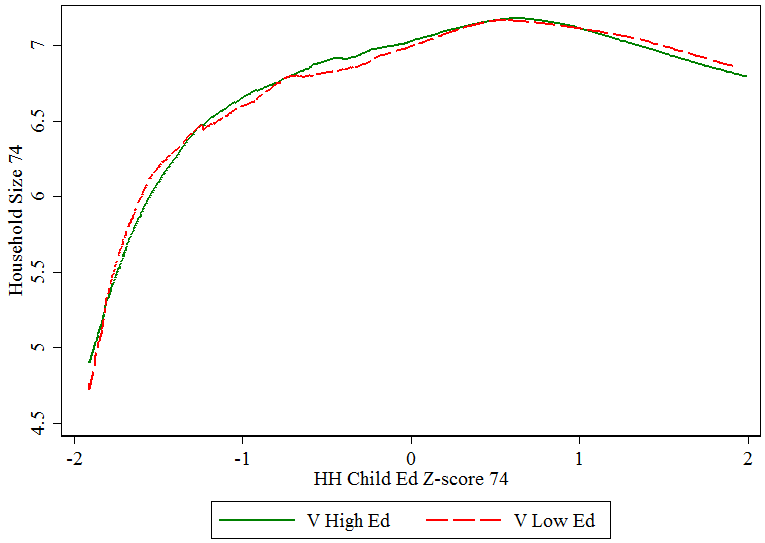
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 7: HH Ed Z-score 2012 by 74 Conditions and 74 Village Ed

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Formal	74 Weights	74/96 Weights	96 Weights	No Weights
Ed Low	-0.0845***	-0.118**	-0.111**	-0.0981***	-0.104***
	(0.0299)	(0.0459)	(0.0464)	(0.0292)	(0.0285)
Ed High	0.100***	0.201***	0.197***	0.198***	0.198***
	(0.0365)	(0.0540)	(0.0544)	(0.0347)	(0.0338)
H Size Low	-0.0434	-0.0692	-0.0614	-0.0213	-0.0259
	(0.0320)	(0.0512)	(0.0521)	(0.0310)	(0.0305)
H Size High	0.000106	-0.00181	-0.00830	0.0501	0.0532
	(0.0337)	(0.0520)	(0.0528)	(0.0385)	(0.0381)
Cons Low	-0.0410	-0.0674	-0.0684	-0.0561*	-0.0566*
	(0.0316)	(0.0464)	(0.0461)	(0.0323)	(0.0317)
Cons High	0.0383	0.105**	0.101*	0.0265	0.0355
	(0.0346)	(0.0529)	(0.0552)	(0.0342)	(0.0336)
V High Ed	-0.00538	-0.0249	-0.0167	0.0289	0.0264
	(0.0272)	(0.0428)	(0.0433)	(0.0282)	(0.0277)
Constant	0.407***	0.733***	0.733***	0.693***	0.689***
	(0.0385)	(0.0546)	(0.0552)	(0.0376)	(0.0371)
Observations	3,646	3,646	3,646	3,646	3,646
R-squared	0.023	0.042	0.038	0.027	0.030

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 10: 74 Household Size by 74 HH Ed Z-score and 74 Village Ed

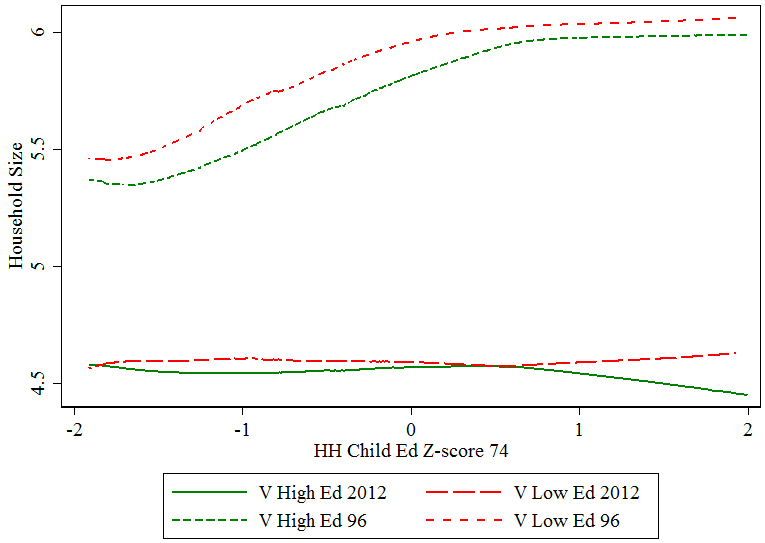


7 Factores contributing to educational mobility

While our primary interest is in educational mobility of households, a key conjecture of this work is that the underlying process of household educational mobility will be related to basic changes in household size and structure. If this is indeed the case then one would expect to see some related changes in these outcomes over time.

To analyze this link between educational mobility and household size change, we are able, in part, to exploit data for the full Matlab population available from the DSS and census data. We first look at household size by 1974 educational investment and village literacy in 1974. Figure 10 examines household size in 1974 and 11 shows household size in 1996 and 2012. It is evident that in 1974 there was little difference by village literacy and that low educational investment households were on average smaller than were high educational investment households. Interestingly, by 1996 households in both strata of village had declined by about 1.5, and this decline was greater in the more literate 1974 villages than the less literate ones. This decrease likely affects the reductions in fertility attributed to the Matlab treatment program in 1978 and the subsequent expansion to the comparison

Figure 11: Household Size by 74 HH Ed Z-score and 74 Village Ed



area in the early 1990s. It is striking, however, that by 2012 the average decline has continued but there is little average difference in household size in the two villages or across 1974 educational strata. The correspondence with the education results is striking and suggests, as might be anticipated given a quality-quantity tradeoff, that expansions in educational investment and reductions in household size go hand in hand.

Of course, there is also a tradeoff between reductions in household size and the expansion in descendants. Other conditions equal, households that have more descendants will on average be smaller than households with fewer descendants. While we do, as one would expect, see in Figure 12 a secular increase in descendants over time, there seems little difference between high and low literacy villages in this regard. Thus the difference in descendants does not seem to help explain the educational investment differences in 1996 and the lack thereof in 2012.

Consumption growth between 1974 and 1996 is plotted in Figure 13. Interestingly, one sees some decline in consumption in the high educational investment households overall, with a greater decline in the higher literacy village. On the other hand the moderate educational investment households in 1974 saw an in-

Figure 12: Descendant Links by 74 Village Ed

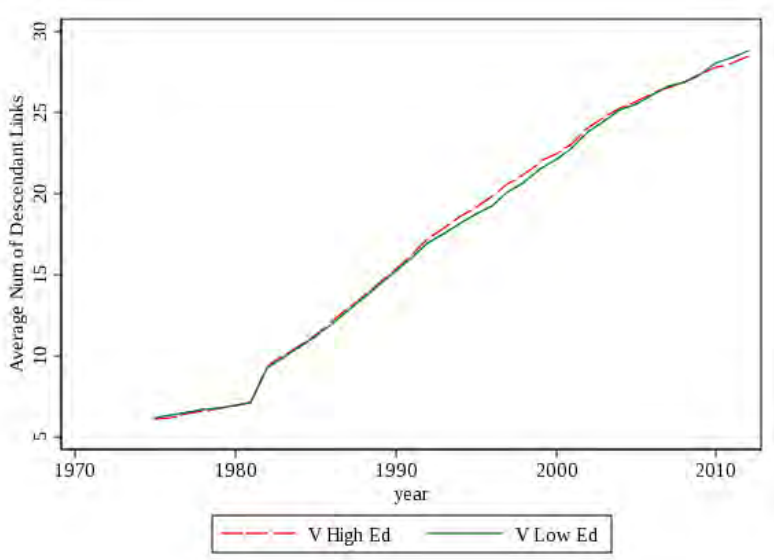


Figure 13: Consumption Growth by 74 HH Ed Z-score and 74 Village Ed

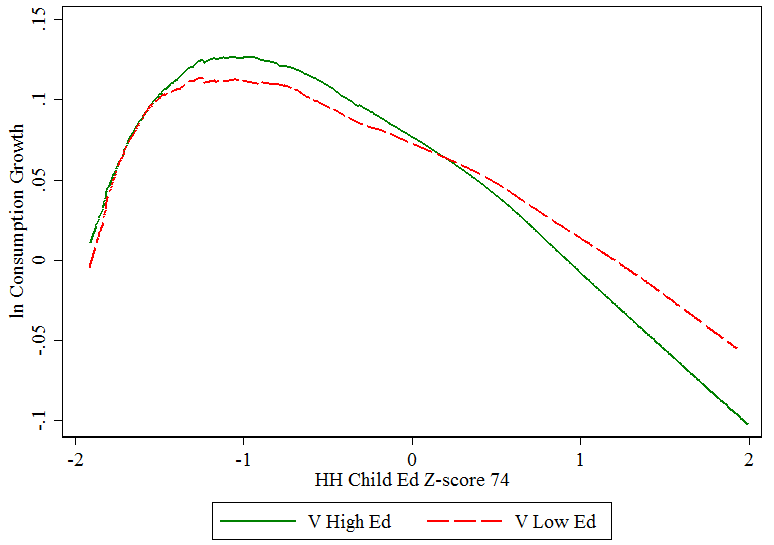
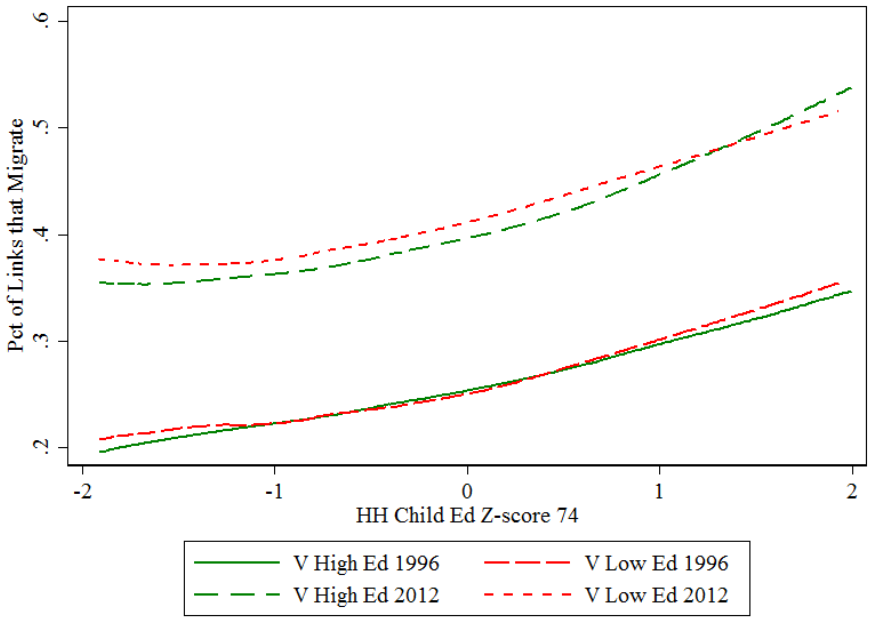


Figure 14: Percent of Links that Migrate by 74 Village Ed



crease in per capital consumption in both types of villages, with a higher increase in the high literacy villages. Differential consumption growth also seems not to play much role in helping to explain educational investment mobility.

Finally Figure 14 looks briefly at migration, another possible source of variation in household size change. These data are based on people who have exited the study area for more than six months and are not present during the relevant survey. Note that exits are indeed substantial, with between 20 and 30 percent of links starting from 1974 being gone by 1996 and between 35 and 52 percent of links being gone by 2012. As might be anticipated, one sees higher outmigration from descendant households in high educational investment lines than low educational investment lines. However, conditional on the level of educational investment in 1974, high literacy villages have lower outmigration. While migration is obviously an important element of mobility and clearly affects the composition and well-being of Matlab residents, it is clear that differential migration cannot explain mobility patterns or the village-level convergence.

Table 8: Population Estimates 1996 by HH Child Z-score and 74 Village Ed

VARIABLES	(1) Change in Household Size	(2) Change in Household Size	(3) Descdant HHs	(4) Descdant HHs	(5) Consumption Growth	(6) Consumption Growth
Ed Low	-0.0523	0.0223	0.197**	0.115**	-0.0318***	-0.0347***
	(0.0529)	(0.0344)	(0.0779)	(0.0520)	(0.00442)	(0.00295)
Ed High	-0.225***	-0.110**	-0.582***	-0.762***	0.0338***	0.0247***
	(0.0714)	(0.0433)	(0.0913)	(0.0556)	(0.00566)	(0.00351)
H Size Low	2.015***	2.046***	-1.234***	-1.186***	-0.0236***	-0.0252***
	(0.0521)	(0.0335)	(0.0737)	(0.0467)	(0.00469)	(0.00312)
H Size High	-3.337***	-3.411***	1.635***	1.625***	0.0582***	0.0634***
	(0.0823)	(0.0526)	(0.117)	(0.0754)	(0.00559)	(0.00353)
Cons Low	-0.577***	-0.597***	0.219***	0.363***	0.166***	0.168***
	(0.0600)	(0.0391)	(0.0847)	(0.0554)	(0.00448)	(0.00296)
Cons High	0.290***	0.303***	-0.318***	-0.311***	-0.225***	-0.232***
	(0.0628)	(0.0394)	(0.0854)	(0.0517)	(0.00564)	(0.00358)
VH x Ed Low	0.132*		-0.141		-0.00449	
	(0.0697)		(0.106)		(0.00590)	
VH x Ed High	0.184**		-0.286**		-0.0143**	
	(0.0899)		(0.113)		(0.00721)	
VH x H Size Low	0.0526		0.0829		-0.00288	
	(0.0676)		(0.0925)		(0.00628)	
VH x H Size High	-0.122		-0.0232		0.00841	
	(0.108)		(0.150)		(0.00726)	
VH x Cons Low	-0.0394		0.259**		0.00382	
	(0.0788)		(0.114)		(0.00590)	
VH x Cons High	0.0184		0.0151		-0.0112	
	(0.0803)		(0.108)		(0.00726)	
V High Ed	-0.260***	-0.181***	0.00461	0.0127	0.0230***	0.0167***
	(0.0740)	(0.0320)	(0.109)	(0.0466)	(0.00585)	(0.00266)
Constant	-0.755***	-0.803***	4.222***	4.218***	0.0802***	0.0841***
	(0.0579)	(0.0418)	(0.0878)	(0.0645)	(0.00455)	(0.00333)
Observations	19,820	19,820	19,822	19,822	19,313	19,313
R-squared	0.461	0.461	0.120	0.120	0.463	0.462

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

These basic bivariate patterns are by and large reflected in multivariate regressions that are presented in tables 8 and 9. We include each of the household variables divided into three categories with the middle category being the reference along with the village literacy variables. We estimate with and without interactions between the village and household level variables. Overall, by 1996, we see higher educational investment in 1974 is associated with lower household size growth, fewer descendants, and higher consumption growth. By contrast larger households are associated with reductions in household size, higher numbers of descendants and higher consumption growth. And higher consumption households in 1974 are associated with more household size growth, fewer descendant households, and lower consumption growth. Finally the high literacy villages show smaller change in household size, no difference in descendants, and greater consumption growth, though overall the effect sizes are quite small. By 2012, the village difference in household size is significant but 1/3 the magnitude and the descendants are marginally higher in the high literacy village.

Table 9: Population Estimates 2012 by HH Child Z-score and 74 Village Ed

VARIABLES	(1) Change in Household Size	(2) Change in Household Size	(3) Descdant HHs	(4) Descdant HHs
Ed Low	0.207***	0.250***	0.0778	-0.0486
	(0.0387)	(0.0256)	(0.161)	(0.108)
Ed High	-0.399***	-0.316***	-1.185***	-1.437***
	(0.0565)	(0.0349)	(0.197)	(0.120)
H Size Low	2.324***	2.342***	-2.646***	-2.455***
	(0.0383)	(0.0246)	(0.150)	(0.0967)
H Size High	-3.606***	-3.673***	3.322***	3.454***
	(0.0670)	(0.0437)	(0.250)	(0.158)
Cons Low	-0.421***	-0.421***	0.321*	0.612***
	(0.0438)	(0.0294)	(0.180)	(0.119)
Cons High	0.270***	0.264***	-0.588***	-0.623***
	(0.0483)	(0.0298)	(0.177)	(0.110)
VH x Ed Low	0.0757		-0.218	
	(0.0516)		(0.217)	
VH x Ed High	0.134*		-0.398	
	(0.0719)		(0.248)	
VH x H Size Low	0.0302		0.327*	
	(0.0498)		(0.194)	
VH x H Size High	-0.113		0.212	
	(0.0886)		(0.325)	
VH x Cons Low	-0.00300		0.516**	
	(0.0585)		(0.240)	
VH x Cons High	-0.0126		-0.0511	
	(0.0612)		(0.227)	
V High Ed	-0.112**	-0.0667***	0.00159	0.162*
	(0.0532)	(0.0246)	(0.228)	(0.0951)
Constant	-2.079***	-2.107***	8.524***	8.428***
	(0.0422)	(0.0307)	(0.177)	(0.128)
Observations	19,175	19,175	19,177	19,177
R-squared	0.633	0.633	0.115	0.114

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

8 Conclusion

We started out wanting to understand economic mobility in a setting in which there had been substantial investment in reproductive health and other primary care services. In order to properly conduct this analysis, though, we had to think through the broad question of how to devise the appropriate weights for panel data when there might be bias in the selection of the sample, and we created those weights based on the MHSS and HDSS data from Matlab, Bangladesh. We first laid out the various issues that arise with the Matlab data due to the post-1978 selection of the MHSS sample. Although this is a problem specific to the Matlab data, it is one that might apply in any of the other HDSS sites where an intervention was conducted on a sample of the population, or only a sample of the HDSS population was later tracked after an intervention in the region. It could arise even in the case of regular panel data if the formation and recombination of households combined with the choice of descendants picked to be surveyed leads to selection bias in the sample that is followed up. Therefore, in the case of any development intervention where there are such data limitations, we have created a possible framework for weights that can help to mitigate the bias.

We devised a procedure to help solve the two main problems with the MHSS/HDSS data. For the first, to make the 1996 sample representative of the 1974 population,

we used the nature of the HDSS data which allowed us to mimic the process that had been used to create the original sample in order to come up with probability weights. Even in the case of panel data where the full population is not available to conduct this sort of resampling procedure, propensity score weights also give extremely good results in helping to correct the sampling bias. For the second problem, we have found a formula for weights that can be universally applied in the case of multiple descendants where not all descendants have the same probability of being picked. Nevertheless, the application of these weights is not advisable if there is no correlation between the probability of being selected and the characteristics of interest, or if the analysis is not focused on aggregate data.

Using the sample weights to look at the main question of interest, we found evidence to support the conclusion that there was a high degree of economic mobility in the area during a period in which access to primary health care services became increasingly available. We do not however see evidence that this change was also associated with a reduction in inequality. At some level, these conclusions seem unsurprising. Access to basic health care can expand opportunity sets both for poor and better off households but this can both expand the variation in outcomes as well as increase the mean absent a deliberate attempt at redistribution through, for example, means tested transfer programs.

References

AMR, Chowdhury et al. (2013). "The Bangladesh paradox: exceptional health achievement despite economic poverty". In: *The Lancet*, online.

Fitzgerald, John, Peter Gottschalk, and Robert A Moffitt (1998). *An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics*.

Foster, Andrew D (1993). "Household partition in rural Bangladesh". In: *Population Studies* 47.1, pp. 97–114.

Foster, Andrew D and Mark R Rosenzweig (2002). "Household division and rural economic growth". In: *Review of Economic Studies*, pp. 839–869.

Joshi, Shareen and T Paul Schultz (2007). *Family planning as an investment in development: evaluation of a program's consequences in Matlab, Bangladesh*. Tech. rep. IZA Discussion Papers.

Moffit, Robert, John Fitzgerald, and Peter Gottschalk (1999). "Sample attrition in panel data: The role of selection on observables". In: *Annales d'Économie et de Statistique*, pp. 129–152.

Pitt, Mark M, Rosenzweig Mark R, and M Nazmul Hassan (2012). "Human Capital Investment and the Gender Division of Labor in a Brawn-Based Economy". In: *American Economic Review*, pp. 3531–60.

Rahman, Omar et al. (1999). "The 1996 Matlab Health and Socioeconomic Survey: Overview and User's Guide". In: *RAND Corporation Draft Series DRU-2018/1-NIA*.

Roy, Nikhil and Andrew D Foster (1996). *The Dynamics of Education and Fertility: Evidence from a Family Planning Experiment*. Tech. rep. University of Pennsylvania.

Schultz, T Paul (2009). "How Does Family Planning Promote Development? Evidence from a Social Experiment in Matlab, Bangladesh 1977-1996". In: *Yale University, Economic Growth Center, New Haven, Conn*.

Sen, A (2013). "What's happening in Bangladesh". In: *The Lancet*, online.