Health and wellbeing of young people in Ethiopia, India, Peru and Vietnam: Life course impacts

Abstract

Using data from four waves of the Young Lives longitudinal survey, we follow the lives of 3,064 eight year old children over 12 years in four developing countries (Ethiopia, India, Peru and Vietnam) to explore the links between children's lives and their health and wellbeing in early adulthood. We apply a novel combination of sequence analysis with clustering and difference-in-differences estimation techniques to identify links between health and wellbeing outcomes in early adulthood and six distinct clusters grouping similar life course pathways. The latter are characterised by family living conditions, economic status and experience of critical life events (including economic shocks). Our results indicate that there were significant differences in health and wellbeing between children in the most advantaged and less advantaged clusters. These wellbeing gaps all narrowed over time but only completely closed for one cluster. In contrast, only some of the initial health gaps narrowed. These results suggest that policy aimed at improving health and wellbeing outcomes in early adulthood needs to focus on supporting disadvantaged young children.

Keywords: Early life poverty, socioeconomic conditions, young adulthood, health, wellbeing, developing countries, sequence analysis.

Word count: 7997 words (excluding title, abstract, tables and figures – as per journal guidelines)

1. Introduction

This paper analyses the effect of early-life socioeconomic conditions on the health and wellbeing of children as they reach early adulthood in four developing countries. The importance of childhood as a determinant of later life outcomes has been highlighted repeatedly over the last decades, and across disciplines from a theoretical as well as empirical perspective (Van De Mheen, Stronks and Mackenbach, 2008; Wellard and Secker, 2017). In particular, recent studies have associated earlier-life conditions with adulthood health and wellbeing (Case and Paxson, 2010; Fors, Lennartsson and Lundberg, 2009), employment prospects (Johnson, Kalil and Dunifon, 2012), and educational attainment (Canning, 2012).

An obstacle in identifying and analysing these effects for developing countries is the limited availability of sufficiently large longitudinal datasets. The Young Lives survey (Boyden, 2016) used in this study is addressing this issue and has to date followed the lives of young children over 12 years in four countries (Ethiopia, India, Peru and Vietnam). This analysis uses a two-step approach to analyse this dataset with the aim of capturing the sequential and interdependent nature of children's lives. First, sequence analysis is used to identify six clusters grouping children who follow distinct life pathways in terms of their access to amenities, schooling and work histories and their experience of economic and family related shocks. Second, regression analysis with difference-in-differences (DID) is used to identify how the six clusters of children are differentiated in terms of their health and wellbeing when first measured and whether or not these differences widen or narrow over time.

Our results indicate that there were significant differences in health and wellbeing between the most advantaged clusters and other, less advantaged groups. While wellbeing gaps narrowed over time, they only completely closed for a cluster of children whose access to amenities significantly improved over their life courses. In contrast, not all of the initially observed health gaps between the clusters narrowed. These results have implications for the later life outcomes of the children who are disadvantaged by their health and wellbeing through childhood and into early adulthood. They support the arguments of Heckman and Masterov (2007) that policy interventions are likely to be more impactful when they focus on aiding young children from disadvantaged environments.

The rest of the paper is structured as follows. Section two provides an overview of the literature on the links between early-life conditions and adult life wellbeing, health and other outcomes. Section three summarises the key features of the dataset and the sample. Section four outlines the methodology and presents the results which are discussed in section five. Section 6 concludes.

2. Research Background

The transition from childhood into adulthood is pivotal in the life course. A number of studies have analysed the different ways in which early life conditions can affect later life outcomes. For instance, early life health indicators, have been associated with adulthood health and welfare outcomes (Bharadwaj, Løken and Neilson, 2013; Conley and Bennet, 2001), learning outcomes (Salm and Schunk, 2008) and future earnings (Black, Devereux and Salvanes, 2007). Similarly, environmental and socioeconomic conditions during early life are often reported to have long-term effects. One example is that of Van Den Berg, Lindeboom and Portrait (2007) who find adult life expectancy to be lower for children born during economic downturns.

Most of the studies that have examined the different pathways to adult life outcomes have however been conducted in the context of developed countries. These studies show that lower family income is associated with negative outcomes such as school dropout, lower educational attainment and health problems (Duncan, Morris, and Rodrigues, 2011). Socioeconomic status of households and overall living conditions have also been found to be related to a range of health problems in adulthood (Melchior, Moffitt, Milne, Poulton and Caspi, 2007).

Early-life environmental and socioeconomic factors are likely to be particularly important for future life outcomes in developing counties, where children are more vulnerable to shocks (such as famine, economic and health shocks). Studies on developing countries have focused primarily on the effect of initial circumstances including birth characteristics, such as height and weight (Vogl, 2012); malnutrition in early life (Field, Robles and Torero, 2009; Linnemayr and Alderman, 2011); famine (Dercon and Porter, 2010); disease (Baird, Hicks, Kremer and Miguel, 2011); environmental pollution (Ferrie, Rolf, and Troesken, 2012; Hanna, Duflo and Greenstone, 2016); and armed conflict (Akresh, Lucchetti and Thirumurthy, 2012). Currie and Vogl (2013) offer a recent and comprehensive survey of this literature.

Other studies focus on single transitions when trying to examine how childhood socioeconomic circumstances affect adult life outcomes (for a review see Juárez and Gayet, 2014). For instance, in Sub-Saharan Africa, the focus has been on individual transitions and outcomes such as dropping out of school (Lloyd and Mensch, 2008), entry into parenthood (Timaeus and Moultrie, 2015), as well as cognitive function (Kobayashi et al, 2017). These transitions and outcomes are often associated with social inequality and poverty that restricts or modifies opportunities for better life outcomes. In a developing country context, these considerations are important given recent efforts to achieve a higher level of educational attainment, wellbeing, and health for all (UN, 2015).

However, focussing on conditions or states at a point in time or single transitions or events ignores the interconnectedness of the processes underlying child development outcomes in developing countries. These can be better understood by adopting a life course approach that recognises the importance of context, changes over time and how the past can shape the future (Holstein and Gubrium, 2000; Kok, 2007; Madero-Cabib, 2015). This is the approach adopted by

Bennett and Waterhouse (2018) who analyse different life course pathways into adulthood among young women in South Africa and self-reported health outcomes.

Our study adopts a similar approach to determine how different childhood histories affect health and wellbeing outcomes later in life. We use the first four rounds of the Young Lives survey data to construct multi-dimensional life-course sequences for 3,064 children in Ethiopia, India, Peru and Vietnam. Analysing the data as sequences that incorporate children's varying life circumstances, rather than focusing on a point in time, allows us to capture the sequential and interconnected character of the different states and processes that underlie childhood development. The sequence analysis identifies six pathways that characterise children's lives in terms of their living conditions, economic status, and experience of family related and economic shocks over their life courses. We then use regression analysis to examine whether these different pathways are linked to health and wellbeing outcomes and trajectories in line with theories of cumulative (dis)advantage (Dannefer, 2003).

3. Data and sample

This paper uses data from the Young Lives project, an international longitudinal cohort study of childhood poverty in four low-middle income countries; Ethiopia, India, Peru and Vietnam. Young Lives selected 12,000 children to be followed, 3,000 in each country from two different cohorts – 2,000 from a younger cohort and 1,000 from an older cohort. Data were collected during four rounds of data collection over 12 years in 2002, 2006, 2009 and 2013. The first round of the survey took place when the younger cohorts were one-year-old and the older cohorts were eight-years-old. The samples are not nationally representative, poorer children were deliberately over-sampled (based on the ranking of regional development indicators) and in India only the states of Andhra Pradesh and Telangana are represented.

In this study, we restrict the analysis to the older cohort because our focus is on outcomes in early adulthood and our constructed life histories capture records of schooling; only the older cohort were entering early adulthood at age 19 in the fourth wave of data collection (the latest wave available at the time of writing) with some already in employment, and only this cohort were in school in the first wave. In round 1, records are accessible for only 3,722 children in this cohort (fewer than the 4,000 targeted) and due to attrition and missed survey years, complete records for all four rounds were only available for 3,095 children. Missing values for some variables further restricted the sample to 3,064 children, comprising 12,256 person-year observations. Attrition is clearly a limitation of the survey data, however analysis by Young Lives (2018) concludes that attrition rates are low compared with other longitudinal data sets in developing countries and are very unlikely to lead to significant biases[1].

The survey collected detailed information on children's individual, family and household characteristics including access to amenities, exposure to external shocks including family and economic shocks and school enrolment. Further details of the survey design and sampling are available from Young Lives (2018) and have been described elsewhere (see, Dendir, 2014; Favara, 2017).

4. Empirical analysis

4.1 Method

The data were constructed as life histories sequences for the 3,064 children and analysed using sequence analysis with optimal matching and clustering (Brzinsky-Fay, Kohler & Luniak, 2006). The resulting clusters were then used in regression analysis.

The life history sequences were constructed through an initial coding process of multiple environmental and socioeconomic factors expected to have significant and longer term impacts on children's health and wellbeing, specifically: (i) living conditions captured by household access to electricity, piped drinking water, their own (not shared) toilet and adequate fuels (paraffin, kerosene, gas or electricity) for cooking; (ii) economic status, recorded by whether or not they are enrolled in school and/or in paid work; and (iii) exposure to unexpected events or shocks that have a negative economic consequence for the home, either family related (divorce, separation, family death or illness) or economic (loss of employment, source of income or family enterprise) since last surveyed, or in the last three years when first surveyed. These three overarching states were coded as follows:

i. Living conditions/access to amenities: Home has access to electricity, own toilet, piped drinking water, adequate fuels for cooking (full access, A); access to 1-3 of these amenities (some access, B); no access to these amenities (C)
ii. Economic status/school enrolment and work: Neither school nor work (M); school with/without work (N); paid work, not in school (0)
iii. Family and economic shocks: no shock (S); family shock only (T); economic shock only or with family shock (U).

The coding results in a variable comprising 27 (3x3x3) possible categories or states recorded sequentially over the survey rounds (supplementary Table S1 provides the distribution for all 27 states). Further sub-categorisation (e.g., by coding the small numbers of those at school and also in work, or those suffering only economic shocks as separate states, or by including states recording other measures of wealth, or shock) would have resulted in states with even fewer observations, potentially weakening the reliability of the clustering procedures (Mojena, 1977).

The coding enables the Stata 13 SQ-Ados scripts (Brzinsky-Fay, et al., 2006) to designate the data as sequence data. In this format, each child's life history is recorded as a sequenced entity simultaneously capturing living conditions, economic status and shocks over 12 years from age eight to nineteen.

4.1.1 Sequence procedures

The multi-dimensional sequences were analysed using optimal matching and cluster analysis (Brzinsky-Fay, et al., 2006). These procedures are used to systematically group similar sequences, treating them as entities and simultaneously taking account of the duration of, and transitions between, different states. The use of optimal matching with cluster analysis to analyse sequenced data is a well-established method and the procedures are well-documented (see for example, Anyadike-Danes and McVicar, 2010). These procedures have been widely used in research on careers (Anyadike-Danes and McVicar, 2010; Wahrendorf, 2015) family formation and work-family trajectories (Aassve, Billari and Piccareta, 2007; Davia and Legazpe, 2014) and in research on life trajectories and health (Bennett and Waterhouse, 2018).

In optimal matching the non-parametric Needleman-Wunsch alignment algorithm is used to compare sequences in order to identify how similar or different they are to each other. This is done by computing the minimum distance, in terms of elementary operations, to turn one sequence into another through substitution or deletion (see in particular Brzinsky-Fay et al*.,* 2006). As positioning within the sequences is important, the insertion/deletion (indel) costs were fixed at half the maximum substitution cost. All the sequences were the same length and therefore there was no requirement for standardisation (Brzinsky-Fay et al., 2006, p. 450). The substitution costs were generated using a data-driven, symmetric transition, frequency-based substitution cost matrix in which non-frequent transitions are costlier (Rohwer & Pötter, 2005). The resulting dissimilarity matrix was used with the widely-used Ward's hierarchical cluster linkage algorithm to group similar sequences in a systematic way, accounting for the incidence of different states, their duration and transitions between them. In this process, hierarchical clustering initially treats each sequence as a separate cluster, the closest two sequences are combined and the process continues until there is only one cluster (Everit, Landau, Leese and Stahl, 2011). In Ward's method, the two clusters that result in the minimum increase in the error-sum-of-squares objective function are joined.

After executing the optimal matching and clustering procedures, the Calinski-Harabasz and Duda/Hart Je(2)/Je(1) index cluster stopping rules (for distance matrices) were implemented in order to determine the number of clusters appropriate for grouping the sequences. The pseudo-*F* index for the Calinski/Harabasz cluster stopping rule suggested the five cluster configuration (F=17.55) representing a step up from the four cluster configuration (F = 16.36). The Duda/Hart Je(2)/Je(1) index was highest (0.9978) and the pseudo T-squared value lowest (T=1.52) for the 6 cluster grouping. A stepped higher pseudo T-squared value for the five and seven cluster groupings (T=4.21 and T = 3.08) indicated a distinct jump in the hierarchical clustering due to cluster separation at this stage suggesting that the six cluster solution was optimal.

In order to choose between the five and six cluster configurations, further consideration of the strength of the clustering was undertaken. This focussed on the cluster fusion values which provide the dissimilarity measure at which clusters are fused or split in the hierarchical cluster structure with larger values indicating more distinct clusters. The fusion values indicated that cluster 1 was very strongly distinguished from the other clusters. Clusters 2 and 3 were least strongly distinguished and the five cluster configuration simply combines these two clusters, the other clusters are not affected (Ward's hierarchical cluster linkage algorithm does not produce reversals). However, guided by the iterative procedure described in Potârcă, Mills and Lesnard (2013) these two clusters were considered sufficiently different and the six cluster configuration was retained.

4.1.2 Regression analysis

Multiple regression analysis is used to investigate how early adulthood health and wellbeing outcomes are impacted by the ways in which children's lives unfold while controlling for a range of other influences. A difference-in-differences, DiD, specification additionally enables quantification of how existing differences in health and wellbeing across the clusters (when the children are first observed) widen, narrow or stay the same (i.e. differ) over time. In the analysis, the covariates capture individual characteristics (age of the child in months and gender) family characteristics (educational attainment of the household head, age of household head, and gender of household head) a regional indicator (urban or rural) and dummy variables recording the country in which the child resides.

The estimating equations are as follows:

$$HEALTH_i = \alpha_{10} + \alpha_{11}LAST\_ROUND + \sum_{j=2}^{n=6} \alpha_{12}CLUSTER_j + \sum_{j=2}^{n=6} \alpha_{13}CLUSTER_j * LAST\_ROUND + \alpha_{14}X_i + \alpha_{15}Z_i + \varepsilon_1 \quad (1)$$

$$WELLBEING_i = \alpha_{20} + \alpha_{21}LAST\_ROUND + \sum_{j=2}^{n=6} \alpha_{22}CLUSTER_j + \sum_{j=2}^{n=6} \alpha_{23}CLUSTER_j * LAST\_ROUND + \alpha_{24}X_i + \alpha_{25}Z_i + \varepsilon_2 \quad (2)$$

The dependent variable in equation (1), $HEALTH_i$, is a subjective measure of health, measured on a scale of 1-3, where 1 is very poor to poor and 3 is good to very good. The second outcome variable, $WELLBEING_i$, is the wellbeing of the child, measured on a scale of 1-9 with a score of 1 indicating poor life satisfaction and 9 higher life satisfaction. This measure was constructed from questions relating to different aspects of the children's lives including opportunities for education, the quality of their housing and support from friends. Subjectively rated health measures are strong predictors of morbidity and mortality (Jylhä, 2009) and have been widely used to provide insights into objective overall health. Self-rated life-satisfaction, happiness and wellbeing measures have also been widely used in research (Blanchflower and Oswald, 2004) and there is a substantial body of evidence that supports the validity of such measures (Wheatley, 2017). Equations (1) and (2) are estimated using OLS and ordered logit. Ordered logit is appropriate since both dependent variables are categorical however the interpretation of the results, particularly for interaction terms, is more straightforward with OLS.

In equations (1) and (2) $CLUSTER_j$ indicates the cluster membership of the child, with cluster 1 being the control group. $LAST\_ROUND$ is a dummy variable that captures whether the observation is for the last available round of the survey ($ROUND_j = 4$) when the children were 19 years old, or the baseline round. $CLUSTER_j * LAST\_ROUND$ variables are interactions. In these specifications, membership of clusters 2 to 6 over the years of the survey is being modelled as a quasi-treatment. The coefficients for $CLUSTER_j$ capture the differences between clusters 2 to 6 and cluster 1 at the baseline. The constant plus the coefficient for LAST_ROUND capture the average outcome for cluster 1 in the last survey round. The coefficients on the interactions capture the average difference-in-differences effects of being in clusters 2 to 6 (relative to cluster 1) between the baseline and the last survey round.

X_i is a vector of individual characteristics including gender and age of the child. Family background characteristics and the regional and country indicator variables are included in the vector Z_i. The latter are included to account for unobservable regional and country level policies and development. See Appendix Table A1 for full definitions of all variables and sample means.

Because of changes to the survey questions as the children grew up, there are some data limitations. In particular, the dependent variable, *HEALTH,* is caregiver-reported in survey years one and two and self-rated in years three and four. As a robustness test, Equation (1) was re-estimated using only the self-reported measure of health making the baseline year the third survey round, 2009, when the children were fifteen. In addition, the wellbeing measure is only available from the second survey round, meaning that the baseline year for the wellbeing outcome is 2006, when the children were twelve years old.

A further consideration in the regression analysis was potential endogeneity of the schooling and work component in the coded sequence states. Indeed, there is a significant body of literature that highlights the existence of a negative correlation between poor child health (particularly pre-school) and favourable early life conditions in developing countries. Glewwe and Miguel (2008) provide discussion of this literature and the common econometric problems encountered when analysing child health in less developed countries.[2] However, the transformations integral to the sequence analysis address this issue. First, the sequence state indicator is a transformed variable that simultaneously captures data from eight different variables and bears little relation to any specific component (see Table S1). More importantly, prior to clustering the data are further transformed into sequenced entities recorded as vectors of ordered states over all survey years and then, through the optimal matching, converted into a matrix of distance measures (as explained in more detail above). Statistical tests for the full dataset confirmed that neither the sequence state variable nor the measure of distance retained in the cluster analysis (Brzinsky-Fay et al., 2006) were statistically related to the categorical indicator of schooling and work status.

4.2. Results

4.2.1 Sequence analysis

Figure 1 presents the 'raw' sequence data visually as time(year)-state distribution plots for the participants. The plots show the sample proportions in each state by year as aggregated views of successive slices of time as the participants age. They illustrate visually how the individual sample members move between the different states over time for the full sample and for each country. For ease of visual interpretation, the graphs combine states with fewer observations. Specifically, states with family shocks and economic shocks are not differentiated as the figures for the latter are quite low. Similarly, some states recording school enrolment (with or without work) or work only are also combined where the figures for the latter was also low (see supplementary Table S1).

The graphs highlight that within the sample there is preponderance of states where the child has some access to amenities and is enrolled in school. The pattern is reflected by the two most common sequences which both record some access to amenities with the child enrolled in school and also sometimes working in all years but either without family or economic shocks (repeated 134 times) or with a family shock in one of the four surveyed years (repeated 70 times).

However, there was still a lot of variation and this is very evident when comparing across the four countries. For example, Peru clearly has the most states with full access to amenities and Ethiopia has the least. Ethiopia also has more states with no access to amenities. The graphs also indicate that over time more children had full access to amenities in all the countries although this was only marginally so in Ethiopia. Unsurprisingly, as the children age, there is a moderate increase in states in which the child is in work only, or neither in work nor school and a corresponding decrease in states where the child is enrolled in school. Ethiopia has the largest share of children not in school or paid work (40.7 percent) and with no access to amenities (69.9 percent), whereas India has the largest share of children in paid work. Among the four countries, India has the highest share of children with full access to amenities (44.6 percent), and the highest in terms of children that experienced no shock. Overall, family shocks were far more common than economic shocks. (See Appendix Table A2 for summary data for the whole sample and by country.)

Figure 1. Time-state Distribution Plots: Full sample and by country

The six clusters identified using the optimal matching and clustering procedures described above capture shared aspects of the underlying histories of the children in the sample. Figure 2 shows the time(year)-state distribution plots for the six clusters. Table 1 shows the distribution of overarching states for each cluster.

Figure 2. Time-state Distribution Plots by Cluster

Table 1. Access to amenities, economic status and incidence of shocks by cluster

Cluster 1 is called the *Better-off & stable* cluster because the cluster members have the highest incidence of full access to amenities, they are mostly enrolled in school (89% of states, see Table 1) with some paid work in later years and suffered relatively few shocks (63% of states are without shocks). In the modal sequence (comprising the most frequent state in each period) children have full access to amenities, are in school and suffer no shocks in every surveyed year (see Table 1).

Cluster 2 is named the *Early transition to adult states* cluster as a relatively high percentage of states record that children are in paid work (20%) or are neither in school nor work (24%) suggesting economic inactivity. Cluster members mostly have some access to amenities and suffer marginally more shocks than cluster 1 (58% of states are without shocks). In the modal sequence, a child would have some access to amenities throughout, face no shocks and move out of school into work in the fourth surveyed year.

Cluster 3 is the smallest grouping. It is called the *Transitioning to better-off* cluster as over the surveyed years the number of states with full access to amenities increases dramatically (from 3% of states to 85%). This cluster also suffers from relatively few shocks (71% of states are without shocks) and most of the time children are enrolled in school (87% of states). The modal sequence is characterised by a shift from some access to amenities to full access in the third period, with children remaining in school and not facing shocks.

Cluster 4 is called *Poor-to-average & some instability.* Its members are the poorest in the sample with 44% of states recording no access to amenities. However, their living standards improved over the surveyed years, with 80% of states recording no access to amenities in the first year compared with only 14% in the last year, and 85% of states instead recording some access to amenities. Their lives were also relatively stable; the incidence of shocks is the same as that of cluster 2 (58% of states are without shocks). However, 27% of states record neither school nor work, suggesting economic inactivity, the highest percentage across the clusters. In the modal sequence children move from no access to amenities to some access in period three and experience a family shock in the second year.

Cluster 5 is called the *Average but unstable* cluster as its members have mostly some access to amenities, but they suffer the most shocks (only 34% of states are without shocks). Most of these shocks are family related. The children in this cluster are mainly enrolled in school (86%). The modal sequence is characterised by some access to amenities, school enrolment and family shocks in every year. **Cluster 6** is the largest cluster. It is called the *Averag*e *& stable* cluster as its members have mainly some but not full access to amenities and they suffer the fewest shocks (79% of states are without shocks). This cluster also records the highest percentage of states where the child is enrolled in school (93%). The modal sequence is characterised by some access to amenities, school enrolment and no shocks.

Prior to exploring how health and wellbeing outcomes varied over time between the clusters, we consider the extent to which the cluster members were differentiated prior to the realisation of any subsequent pathway specific effects. Table 2 presents mean differences by cluster for a range of individual, household and locational characteristics in the baseline year.

The figures show that there is a higher representation of females in cluster 3 (*Transitioning to better-off)* compared with the rest of the sample. Cluster 2 (*Early transition to adult states)* are unsurprisingly the oldest cluster while cluster 4 (*Poor-to-average & some instability)* are younger. From the age of 8, children in cluster 1 (*Better-off & stable)* were rated by their main caregiver as relatively more healthy than their peers while children in cluster 2 (*Early transition to adult states)* were rated as less healthy. Children in cluster 1 also had higher wellbeing when they were 12 as did those in cluster 6 (*Average and stable)*. Children in clusters 2 and 5 (*Average but unstable)* had lower wellbeing.

Table 2. Cluster characteristics (mean differences)

Head of household educational attainment is, perhaps unsurprisingly, highest in cluster 1 (*Better-off & stable)* and lowest in cluster 4, the poorest cluster, cluster 2, the oldest cluster, and cluster 5, the least stable cluster. It may be that educational attainment is related negatively to parents' age as the head of household is also older in clusters 4 and 5 but younger in cluster 6 (*Average & stable),* where the head of household is also most likely to be male.

Children in clusters 1 (*Better-off & stable)* and 3 (*Transitioning to better-off)* were more likely to be living in urban areas, those in 2, 4 and 6 were more likely to be living in rural areas. Children in cluster 5 (*Average but unstable)* were no more or less likely to be living in urban or rural areas. Children in Ethiopia were more likely to be in clusters 4 and 5 (the poorest and the least stable clusters) while, in a complete mirror image, children in India were more likely to be in clusters 1-3. Children in Peru were most likely to be in clusters 1 and 3 (the better off and more urban clusters) but least likely to be in cluster 6 (the cluster suffering the least family and economic shocks). Children in Vietnam were most likely to be in clusters 2 and 6, the early transitioning and most stable clusters respectively).

These results indicate that earlier on in their lives the children in cluster 1 (*Better-off & stable)* were the most advantaged in terms of their health and wellbeing. Children in cluster 6 (the most stable cluster) also scored higher on the wellbeing measure. Children in cluster 2 (*Early transition to adult states*) and cluster 5 (*Average but unstable*) appear to have been the most disadvantaged. The analysis in the next section controls for other influences and also considers whether these differences widened, narrowed or stayed the same over the course of the children's lives.

4.2.2 Regression analysis

Table 3 shows the results of the OLS and ordered logit DiD estimations with the measures of health and wellbeing as dependent variables. For the ordered logits odds ratios are presented.

The average outcome for the reference cluster, cluster 1 (*Better-off & stable)* in the baseline year is captured by the constant (equivalently the cutpoints in the ordered logit). The cluster coefficients (odds ratios) reflect the differences between clusters 2 to 6 and cluster 1 at the baseline ($ROUND_j = 1 \; or \; 2$). The coefficients on the interaction terms capture the difference in these differences between the baseline and the last round, whether they widen, narrow or remain the same. However, the interpretation of the interactions terms in the ordered logit is complicated by the non-linearity of the maximum likelihood estimates as the effect of a change in a variable depends on the values of all the variables in the estimation. This is a particular problem for linked, interacted variables, and interpretation is usually based on predictions for specified values of these variables. To address this, we computed the marginal changes in the probabilities for each cluster at the baseline and final round and used these to calculate the difference-in-differences over time (supplementary Table S2 presents a sub-set of these calculations). Because of this complication and since the pattern of significance of the included variables is almost identical across the OLS and ordered logit estimations, the interpretation of the results focuses on the OLS estimates.

In all estimations, the significant influences of the cluster variables are negative relative to the reference group, cluster 1 (*Better-off and stable*) which reported the best health and wellbeing (see Table 2). In the health estimations, these negative effects are significant for clusters 2, 4 and 5. They are absolutely larger for cluster 4 (*Poor-to-average & some instability*) indicating that when the children were eight, those in cluster 4 recorded the worst health relative to cluster 1. This is consistent with cluster 4 being the poorest cluster in terms of access to amenities (Table 1). The lack of significance of clusters 3 (*Transitioning to better-off*) and 6 (*Average & stable*) indicates that in the baseline year children in these two clusters did not have significantly different health from those in cluster 1. This is consistent with cluster 3 being the second most well-off cluster in terms of living conditions and cluster 6 the most stable (Table 1).

Table 3. Regression estimates for health and wellbeing

The DiD analysis is captured by the interaction effects between the cluster variables and the last survey round indicator. In the health estimates, the significant and positive coefficient (odds ratios greater than 1) on the interaction term for cluster 4 (*Poor-to-average & some instability*) indicates that the health gap between cluster 4 and 1 narrowed over time. Furthermore, a comparison of the OLS coefficients of the cluster variables and the interaction terms shows that this gap not only narrowed but also closed. Equivalently, in the ordered logit estimates, the marginal change in the probability of reporting good/very good health associated with cluster 4 membership at the baseline is -0.0895, rising to 0.0150 in the final round, a difference in the difference of 0.0823 (supplementary Table S1). The health gap for cluster 4 was initially the largest and the relative improvement seems likely to be linked to the parallel improvement in living conditions. A similar improvement in living conditions was not experienced in clusters 2 (*Early transition to adult states*) or 5 (*Average but unstable*) and the corresponding health gaps did not close. The insignificant baseline cluster effects for clusters 3 (*Transitioning to better-off*) and 6 (*Average & stable)* are matched in the OLS estimates by insignificant interaction terms indicating continued parity with cluster 1. However, the negative interaction term for cluster 3 is just weakly significant in the logit suggesting that the health of children in cluster 3 may have fallen behind (as shown in supplementary Table S2 the difference in the difference is -0.0804). This weak effect is difficult to explain although may be related to higher female representation.

In the wellbeing estimations, all the cluster effects are negative and significant indicating that children in cluster 1 (*Better-off & stable*) reported higher wellbeing. These differences are absolutely largest for cluster 2 (*Early transition to adult state*s) cluster 4 (*Poor-to-average & some instability*) and cluster 5 (*Average but unstable*). Similar to the results for health, wellbeing gaps are smaller for clusters 6 (*Average & stable*) and 3 (*Transitioning to better-off*).

The interaction terms in the wellbeing estimations are all significant and positive indicating that all the wellbeing gaps between cluster 1 and the other clusters narrowed between 2006 and 2013. The OLS coefficients indicate that the proportionate reduction in wellbeing gaps was greatest for clusters 3 (*Transitioning to better-off*) and 6 (*Average & stable*). However, the wellbeing gap only completely closed for cluster 3: in the logit, the marginal change in the probability of reporting the mean value for wellbeing is -0.0112 for cluster 3 at the baseline and 0.0004 in the final round, a difference in the difference of 0.0211 (supplementary Table S3). The dramatic improvement in living conditions in cluster 3 is likely to be a determining factor underlying this result.

The results for the other included variables are also interesting. Female children reported significantly worse health but higher wellbeing. Children in families with a more educated head of household had better health and wellbeing while male and older household heads are associated with higher wellbeing. Children living in rural locations had no better or worse health than those living in urban locations, but they recorded higher levels of wellbeing. After controlling for life histories through cluster membership and individual and family characteristics, children in Vietnam and Peru recorded the lowest subjective health but children in Peru also reported the highest wellbeing while children in India reported the lowest wellbeing.

As test of robustness, we replaced the dependent variable in equation (1) with self-rated health available only in rounds three and four. The directions of the effects and the patterns of significance of the cluster variables and most interaction terms remained consistent (Supplementary Table S3). The only significant exceptions were weak positive significance of the cluster 5 interaction term and consistent insignificance of the cluster 3 interaction. We further experimented by including caregiver reported health in survey year 1 as a control for initial health. This made minimal difference to the results and the variable itself was not significant (not reported).

5. Discussion

The research makes significant empirical and methodological contributions to the literature analysing the transition to adulthood in developing countries. It does this by examining the relationships between young life histories and health and wellbeing outcomes in early adulthood for four developing countries. These relationships are examined using sequence and regression analysis to identify six clusters that captured different life trajectories over twelve years between 2002, when the children in the sample were eight years old, and 2013 when they were nineteen. Our results indicate that many of the links that have been documented in the literature between upbringing conditions and later life health and wellbeing outcomes in developed economies

(Kobayashi et al., 2017) are just as important when considered in the context of developing, and overall poorer, countries.

Within each of the six clusters identified in the analysis, children experienced similar pathways in terms of their living conditions, schooling and work participation and key life events in terms of family and economic shocks. Cluster 1 (*Better-off & stable*) identifies a group of children whose homes had good access to basic amenities (electricity, own toilet, piped drinking water and adequate fuels for cooking) and relatively infrequent family or economic shocks. Children in cluster 2 (*Early transition to adult states*) were the most likely to be in paid work. In cluster 3 (*Transitioning to better-off*) children's access to amenities improved dramatically over time and females were strongly represented. Children in cluster 4 (*Poor-to-average & some instability)* were the poorest, the incidence of shocks was above average but their access to amenities improved over time. Children in cluster 5 (*Average but unstable*) and cluster 6 (*Average & stable*) had mainly some but not full access to amenities but cluster 5 suffered considerably more frequent shocks, particularly family shocks (i.e. marital break-up, a family illness or death). These six clusters reflect the trajectories of our sample but are likely to exist amongst a wider population of young people in developing countries. This can be tested in further research.

The regression results indicate that there are significant links between children's lives, including their experience of key life events or shocks, and their health and wellbeing in their early years and early adulthood. The results indicated that relative to the most advantaged cluster in terms of access to amenities, cluster 1 (*Better-off and stable*) three of the clusters had significantly worse health relative to cluster 1 and although these health gaps narrowed over time for the children in poorest cluster, cluster 4 (*Poor-to-average & some instability*) this was less true for the cluster suffering the most shocks, cluster 5 (*Average but unstable*) and cluster 2 (*Early* *transition to adult states*). The latter result may have some parallels with those of Bennett and Waterhouse (2018) who find that early motherhood and economic inactivity are associated with poorer self-rated health. Interestingly there were no significant health gaps between children in cluster 1 and clusters 3 (*Transitioning to better-off*) and 6 (*Average & stable*) who were relatively advantaged in terms of living conditions and exposure to shocks, respectively. In contrast, there were baseline wellbeing gaps between cluster 1 and all the other clusters but all of these gaps narrowed over time. However, the wellbeing gaps only closed completely for cluster 3. Nevertheless, while not all these health and wellbeing gaps closed, none widened over time suggesting that there were no cumulative disadvantages for the children in the sample.

There are some limitations to this study which suggest avenues for further research. Firstly, the longitudinal data while spanning twelve years is derived from only four survey years which naturally limits the degree of detail and variability in children's lives that is captured by the data. Nevertheless, the longitudinal data enables the construction of multidimensional life course records that represent an improvement on data collected retrospectively. Secondly, limitations on the availability of the measures of health and wellbeing constrain the regression analysis. Third, sequence analysis procedures are limited by the need for researchers to make choices about costs used to calculate the distance matrix and the number of clusters (Halpin, 2010; Potârcă et al., 2013). Further research could build upon our work using longitudinal data collected more frequently over children's lives or for more years. Future research will also benefit from additional rounds of the Young Lives survey which will enable an extended analysis of the life histories of both cohorts. This would enable more detailed consideration of how children's life histories are linked to health and wellbeing outcomes in early adulthood.

6. Conclusion

The analysis of this study has implications for policy focused on reducing inequalities in health and wellbeing for young people in developing countries. First, the results suggest that living conditions, key life events and schooling and work decisions throughout children's life courses are linked systematically to health and wellbeing outcomes in early adulthood. As such, the research highlights the importance of the way children's lives unfold and their experience of family and economic shocks in a developing country context. Second, the results highlight how some initial inequalities in children's health and wellbeing can narrow even while others persist. In particular, wellbeing gaps narrowed the most for children whose living conditions improved significantly (cluster 3, *Transitioning to better-off*) and for those who experienced the greatest stability (cluster 6, *Average & stable*). One implication is that policy interventions designed to support the health and wellbeing of young adults in developing countries need to focus on mitigating family and economic instability as well as improving children's material living conditions. As such, this is consistent with the evidence of Heckman and Masterov (2007) who argue that interventions in early life are likely to promote overall outcomes later on.

Endnotes

[1] Notes to Table A1 include summary analysis of sample attrition.
[2] These include the limited range of variables which can lead to omitted variable bias, and measurement errors which can raise concerns of attenuation bias. The use of panel data sets can alleviate some of these problems (Glewwe and Miguel, 2008).

References

Aassve, A, Billari F. C, Piccareta, R. F. 2007. Strings of adulthood: A sequence analysis of young women's work-family trajectories, *Journal of European Social Policy* 23(3-4), 369-388

Akresh, R., Lucchetti L. & Thirumurthy, H. 2012. Wars and child health: evidence from the Eritrean-Ethiopian conflict. *Journal of Development Economics*, 99, 330–40

Anyadike-Danes, M. & McVicar, D. 2010. My Brilliant Career: Characterizing the Early Labor Market Trajectories of British Women from Generation X, *Sociological Methods Research* 38(3), 482-512.

Baird S, Hicks J. H, Kremer M, Miguel E. 2011. Worms at work: long-run impacts of child health gains, Working Paper*, University of California*, Berkeley.

Black, S.E., Devereux, P.J. and Salvanes, K.G., 2007. From the cradle to the labor market? The effect of birth weight on adult outcomes. *The Quarterly Journal of Economics*, 122(1), 409-439.

Bennett, R. and Waterhouse, P. 2018. Work and Family Transitions and the self-rated health of young women in South Africa, *Social Science and Medicine*, 203, 9-18.

Bharadwaj P, Løken KV, Neilson C. 2013. Early life health interventions and academic achievement. *American Economic Review*. 103(5), 1862–91.

Blanchflower, D., Oswald, A.J., 2004. Well-being over time in Britain and the USA. *J. Public Economics*. 88,1359–1386.

Boyden, J. 2016. *Young Lives: an International Study of Childhood* Poverty: *Rounds 1-4 Constructed Files*, 2002-2014, 2nd Edition, UK Data Service. SN: 7483, http://doi.org/10.5255/UKDA-SN-7483-2

Brzinsky-Fay, C., Kohler U. & Luniak, M. 2006. Sequence analysis with Stata**,** *The Stata Journal* 6(4), 435–460.

Canning, D. 2012. Progress in Health around the World. *Journal of Development Studies*, 48:12, 1784-1798.

Case A, Paxson C. 2010. Causes and consequences of early-life health. *Demography,* 47:S65–S85.

Conley D, Bennett N. G. 2001. Birth weight and income: interactions across generations. Journal of *Health and Social Behavior*, 42, 450–465.

Currie, J., & Vogl, T. 2013. Early-Life Health and Adult Circumstance in Developing Countries, *Annual Review of Economics*, 5 (1):1–36 [425,444]

Dannefer, D. 2003. Cumulative advantage/disadvantage and the life course: cross fertilizing age and social science theory, *Journal of Gerontology: Social Sciences*, 58B(6), S327-337.

Davia, M. A. & Legazpe, N. 2014. Female employment and fertility trajectories in Spain: an optimal matching analysis, *Work, Employment and Society 28*(4), *633-65.*

Dendir, S. 2014. Children's cognitive ability, schooling and work: Evidence from Ethiopia. *International Journal of Educational Development,* 38, 22-36.

Dercon S, Porter C. 2010. Live Aid revisited: long-term impacts of the 1984 Ethiopian famine on children, Working Paper 2010-39, *Centre for Studies in African Economics*, Oxford University.

Duncan, G. J., Morris, P. A., & Rodrigues, C. 2011. Does money really matter? Estimating impacts of family income on young children's achievement with data from random-assignment experiments. *Developmental Psychology*. doi:10.1037/a0023875.

Everitt, B. S., Landau, S.. Leese M, and Stahl. D. 2011. *Cluster Analysis*. 5th ed. Chichester, UK: Wiley.

Favara, M. 2017. Do dreams come true? Aspirations and educational attainments of Ethiopian boys and girls. *Journal of African Economies*, 26(5), 561-583.

Ferrie J, Rolf K, Troesken W. 2012. Cognitive disparities, lead plumbing, and water chemistry: Prior exposure to water-borne lead and intelligence scores among World War Two U.S. Army enlistees. *Economics & Human Biology,* 10(1), 98-111.

Field, E., Robles, O. and Torero, M., 2009. Iodine deficiency and schooling attainment in Tanzania. *American Economic Journal: Applied Economics*, 1(4), 140-69.

Fors, S., Lennartsson, C., & Lundberg, O. 2009. Childhood living conditions, socioeconomic position in adulthood, and cognition in later life — exploring the associations. *Journal of Gerontology: Social Sciences*, 64B(6), 750–757.

Glewwe, P. and Miguel, E.A., 2007. The impact of child health and nutrition on education in less developed countries in Schultz, T. P and John A. Strauss, J. A (eds.) *Handbook of development economics*, 4, 3561-3606, Amsterdam, Netherlands: Elsevier.

Halpin, B. 2010. Optimal Matching Analysis and Life-Course Data: The Importance of Duration. *Sociological Methods & Research*, 38(3), 365-88.

Hanna R, Duflo E, Greenstone M. 2016. Up in smoke: the influence of household behavior on the long-run impact of improved cooking stoves, *American Economic Journal: Economic Policy, ,* 8(1), 80-114.

Heckman J, Masterov D. 2007. The productivity argument for investing in young children. *Applied Economic Perspectives and Policy*, 29, 446–493.

Holstein, J.A. & Gubrium, J.F. 2000. *Constructing the Life Course*. 2nd ed. Dix Hills, New York: General Hall.

Juárez, F., & Gayet, C. 2014. Transitions to adulthood in developing countries. *Annual Review of Sociology*, *40*, 521-538.

Jylhä, M., 2009. What is self-rated health and why does it predict mortality? Towards a unified conceptual model. *Social Science & Medicine*, 69, 307–316.

Johnson RC, Kalil A, Dunifon R. E. 2012. Employment patterns of less-skilled workers: Links to children's behavior and academic progress. *Demography.* 49, 747–772.

Kobayashi, L. C., Glymour, M. M., Kahn, K., Payne, C. F., Wagner, R. G., Montana, L., ... & Berkman, L. F. 2017. Childhood deprivation and later-life cognitive function in a population-based study of older rural South Africans. *Social Science & Medicine*, *190*, 20-28.

Kok, J. 2007. Principles and prospects of the life course paradigm. *Annales du Démographie Historique* 113(1), 203-23.

Linnemayr, S. & Alderman H. 2011. Almost random: evaluating a large-scale randomized nutrition program in the presence of crossover. *Journal of Development. Economics,* 96:106–14

Lloyd CB, Mensch BS. 2008. Marriage and childbirth as factors in dropping out from school: an analysis of DHS data from sub-Saharan Africa. *Population Studies,* 62(1), 1–13.

Madero-Cabib, I. 2015. The life course determinants of vulnerability in late careers, *Longitudinal and Life Course Studies*, 6, 88-106.

Melchior, M., Moffitt, T. E., Milne, B. J., Poulton, R., & Caspi, A. 2007. Why do children from socioeconomically disadvantaged families suffer from poor health when they reach adulthood? A life-course study. *American Journal of Epidemiology*, 166, 966-974.

Mojena, R. 1977. Hierarchical grouping methods and stopping rules: An evaluation, *The Computer Journal*, 20(4), 359-363.

Rohwer, G., and Pötter. U. 2005. Data-based Substitution Costs, in *TDA User Manual* (6.7.2.5 p.1). Bochum, Germany: Ruhr-Universitat Bochum. Retrieved from http://www.stat.ruhr-uni-bochum.de/pub/tda/doc/tman63/tman-pdf.zip. [Accessed 4/06/18]

Salm, M., and Schunk, D. 2008. The role of childhood health for the intergenerational transmission of human capital: Evidence from administrative data. *Institute for the Study of Labor (IZA)*, Discussion Paper 3646

Timæus, I. M., & Moultrie, T. A. 2015. Teenage childbearing and educational attainment in South Africa. *Studies in family planning*, *46*(2), 143-160.

Van Den Berg, G. J., Lindeboom, M., Portrait, F., 2006. Economic conditions early in life and individual mortality. *American Economic Review* 96(1), 290–302.

Van De Mheen, H. D., Stronks, K. and Mackenbach, J. P. 1998. A Lifecourse Perspective on Socio-economic Inequalities in Health: The Influence of Childhood Socio-economic Conditions and Selection Processes, *Sociology of Health & Illness*, 20, 754-777.

Vogl, T.S., 2014. Height, skills, and labor market outcomes in Mexico. *Journal of Development Economics*, 107, 84-96.

Wahrendorf, M., 2015. Previous employment histories and quality of life in older ages: sequence analysis using SHARELIFE. *Ageing & Society* 35(9), 1928–1959.

Wellard, I., & Secker, M. 2017. 'Visions' for children's health and wellbeing: exploring the complex and arbitrary processes of putting theory into practice. *Sport, Education and Society*, *22*(5), 586-601.

Wheatley, D. 2017. *Time Well Spent: Subjective Well-being and the Organization of Time*, London: Rowman and Littlefield International

Young Lives 2018. *Survey Design and Sampling Fact sheets*, www.younglives.org.uk [Accessed 4/06/18]

TABLES

Table 1. Access to amenities, economic status and incidence of shocks by cluster (%)

	Cluster					
	1 *Better-off & stable*	2 *Early transition to adult states*	3 *Transitioning to better-off*	4 *Poor-to-average & some instability*	5 *Average but unstable*	6 *Average & stable*
Living conditions/Access to amenities (% within clusters)						
Full access	84	1	40	1	3	2
Only some access	16	94	59	55	92	96
No access	0	5	1	44	5	2
Economic status: School and/or paid work (% within clusters)						
No school or paid work	4	24	8	27	10	5
School with/without paid work	89	56	87	66	86	93
Paid work only	5	20	6	7	4	2
Shocks (% within cluster)						
No shock	63	58	71	58	34	79
Family shock only	28	37	23	38	52	16
Economic shock only	6	3	3	2	7	3
Economic and family shock	3	2	2	2	7	2
N	1,836	2,756	1,008	1,340	2,140	3,176
Number of sequences (sequence size)	459	689	252	335	535	794
Modal sequence*	(ANS;4)	(BNS;3) (BOS;1)	(BNS;2) (ANS;2)	(CMS;1) (CNT:1) (BNS;1) (BMS;1)	(BNT:4)	(BNS;4)

* See section 4.1 and Supplementary Table S1 for key

Table 2: Cluster characteristics (mean differences in baseline year)

	Cluster 1 *Better-off & stable*	Cluster 2 *Early transition to adult states*	Cluster 3 *Transitioning to better-off*	Cluster 4 *Poor-to-average & some instability*	Cluster 5 *Average but unstable*	Cluster 6 *Average & stable*
Individual characteristics						
Female	-0.0368	0.0117	0.0680**	-0.0174	0.003	-0.006
	(-1.45)	(0.54)	(2.07)	(-0.60)	(0.11)	(-0.29)
Age (in months)	0.115	0.603***	0.181	-1.003***	-0.186	-0.0465
	(0.61)	(3.74)	(0.74)	(-4.66)	(-1.05)	(-0.30)
Health	0.131***	-0.121***	0.0686	-0.059	-0.003	0.0282
	(3.75)	(-4.05)	(1.51)	(-1.48)	(-0.09)	(0.99)
Wellbeing	0.816***	-0.638***	0.204	-0.136	-0.230**	0.196**
	(8.54)	(-7.77)	(1.63)	(-1.23)	(-2.54)	(2.49)
Household characteristics						
Education of household head	4.924***	-2.414***	1.909***	-3.141***	-0.576**	0.160
	(21.38)	(-11.67)	(5.97)	(-11.17)	(-2.47)	(0.79)
Age of household head	0.164	0.527	-0.813	1.708***	1.293**	-2.103***
	(0.30)	(1.13)	(-1.15)	(2.74)	(2.52)	(-4.74)
Female household head	0.0047	-0.011	0.0268	0.008	0.0255	-0.0267*
	(0.25)	(-0.70)	(1.09)	(0.38)	(1.44)	(-1.74)
Locational factors						
Rural location	-0.720***	0.250***	-0.198***	0.327***	-0.0126	0.172***
	(-36.55)	(12.72)	(-6.50)	(12.44)	(-0.57)	(9.05)
Ethiopia	-0.330***	-0.111***	-0.225***	0.475***	0.285***	-0.0463**
	(-14.82)	(-5.67)	(-7.58)	(19.03)	(13.53)	(-2.47)
India	0.169***	0.0846***	0.136***	-0.275***	-0.157***	0.0151
	(7.29)	(4.25)	(4.51)	(-10.47)	(-7.22)	(0.79)
Peru	0.302***	-0.060***	0.0748***	-0.072***	-0.075***	-0.083***
	(19.64)	(-4.33)	(3.54)	(-3.86)	(-4.89)	(-6.25)
Vietnam	-0.141***	0.0869***	0.0138	-0.128***	-0.0529**	0.114***
	(-6.23)	(4.48)	(0.47)	(-4.93)	(-2.47)	(6.17)
Number of sequences =3,064						

Reported figures are mean differences for the sub-sample (mean cluster_j –mean all clusters excluding cluster_j).
t statistics and significance levels reported for t tests of differences in means between cluster_j (j = 1, 6) and the rest of the sample, t statistics in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Statistics for Wellbeing are for survey data round 2 as the measure of wellbeing was not available in round 1.

Table 3. Regression estimates for health and wellbeing

Independent variable	(1a) OLS Health (β)	(1b) O. Logit Health (e^{β})	(2a) OLS Wellbeing (β)	(2b) O. Logit Wellbeing (e^{β})
LAST_ROUND	0.418	7.832**	0.713*	2.285*
	(0.277)	(7.243)	(0.427)	(1.080)
Cluster 2 *Early transition to adult states*	-0.122***	0.701**	-0.965***	0.323***
	(0.0443)	(0.100)	(0.115)	(0.0416)
Cluster 3 *Transitioning to better-off*	0.0116	1.070	-0.378***	0.667***
	(0.0504)	(0.174)	(0.130)	(0.0967)
Cluster 4 *Poor-to-average & some instability*	-0.157***	0.602***	-0.899***	0.343***
	(0.0561)	(0.110)	(0.150)	(0.0584)
Cluster 5 *Average but unstable*	-0.0980**	0.754*	-0.782***	0.394***
	(0.0461)	(0.113)	(0.119)	(0.0532)
Cluster 6 *Average & stable*	-0.0171	0.955	-0.456***	0.563***
	(0.0426)	(0.132)	(0.110)	(0.0697)
LAST_ROUND*Cluster 2 *Early transition to adult states*	0.0300	1.038	0.248*	1.357*
	(0.0546)	(0.190)	(0.143)	(0.215)
LAST_ROUND*Cluster 3 *Transitioning to better-off*	-0.115	0.650*	0.416**	1.507**
	(0.0699)	(0.152)	(0.182)	(0.301)
LAST_ROUND*Cluster 4 *Poor-to-average & some instability*	0.171**	1.815**	0.326*	1.521**
	(0.0671)	(0.423)	(0.180)	(0.304)
LAST_ROUND*Cluster 5 *Average but unstable*	0.0324	1.052	0.336**	1.529**
	(0.0575)	(0.206)	(0.152)	(0.257)
LAST_ROUND*Cluster 6 *Average & stable*	-0.0182	0.902	0.294**	1.441**
	(0.0526)	(0.160)	(0.138)	(0.218)
Female	-0.0562***	0.840***	0.102**	1.138***
	(0.0166)	(0.0465)	(0.0439)	(0.0554)
Age (in months)	-0.00284	0.938	-0.0311	0.967
	(0.0249)	(0.0780)	(0.0627)	(0.0671)
Education of household head	0.00910***	1.033***	0.0669***	1.082***
	(0.00223)	(0.00780)	(0.00594)	(0.00720)
Female household head	-0.0222	0.929	-0.218***	0.778***
	(0.0224)	(0.0699)	(0.0580)	(0.0502)
Age of household head	-0.000005	1.000	0.0107***	1.014***
	(0.000785)	(0.00265)	(0.00212)	(0.00241)
Rural location	0.00158	1.020	0.490***	1.730***
	(0.0242)	(0.0840)	(0.0636)	(0.123)
India	-0.0198	0.899	-0.686***	0.454***
	(0.0266)	(0.0840)	(0.0729)	(0.0374)
Peru	-0.217***	0.424***	0.878***	2.582***

	(0.0323)	(0.0462)	(0.0847)	(0.254)
Vietnam	-0.471***	0.193***	-0.0178	0.980
	(0.0262)	(0.0174)	(0.0708)	(0.0774)
Constant/Cutpoint 1	2.407***	-3.017	4.356***	-3.344
	(0.206)	(0.688)	(0.786)	(0.8746)
Observations	5,396	5,396	5,327	5,327
R^2 /Pseudo R^2	0.179	0.110	0.209	0.063
F/LR Chi^2	60.06***	1141.51***	69.95	1333.33

Reported figures are coefficients (β) for OLS estimates and odds ratios (e^{β}) for ordered logits. Only cutpoint 1 value reported for ordered logits. Standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

FIGURES

Figure 1: *Time-state Distribution Plots: Full sample and by country*

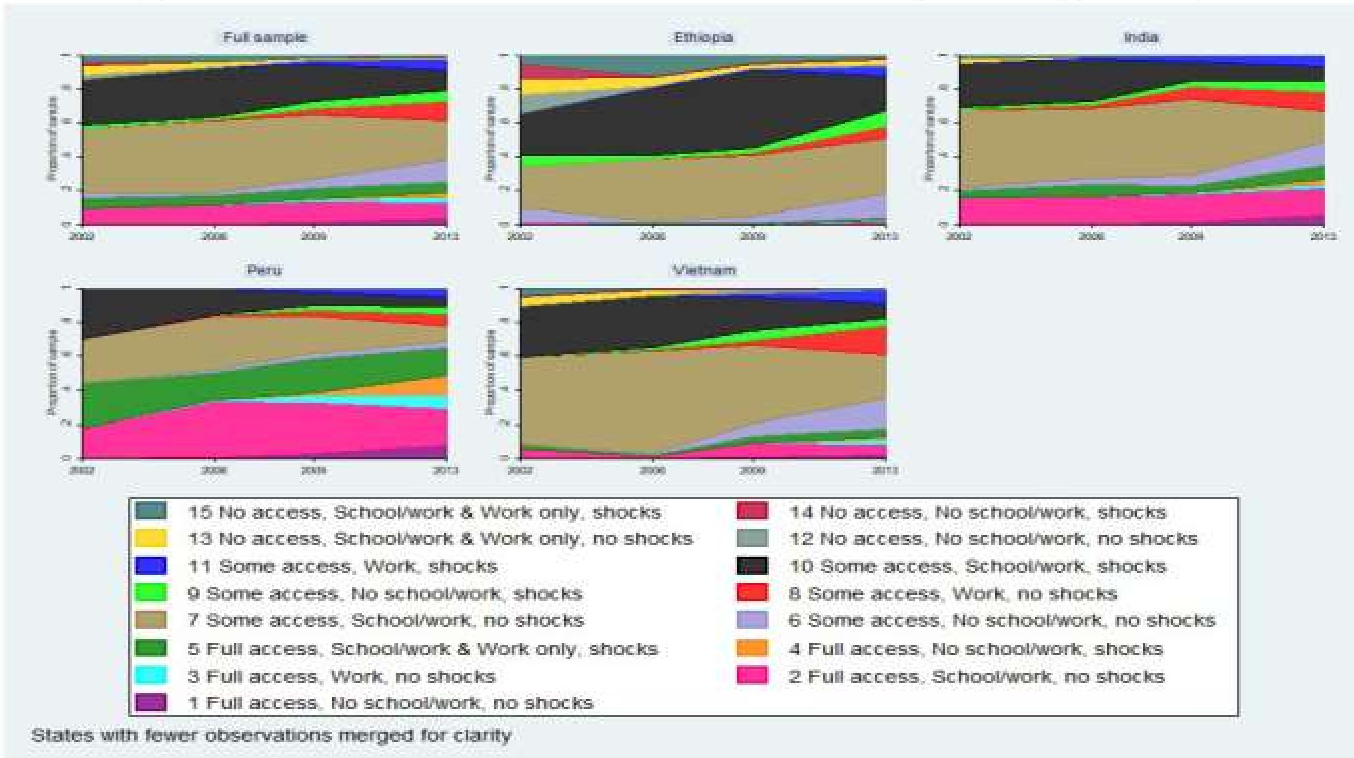
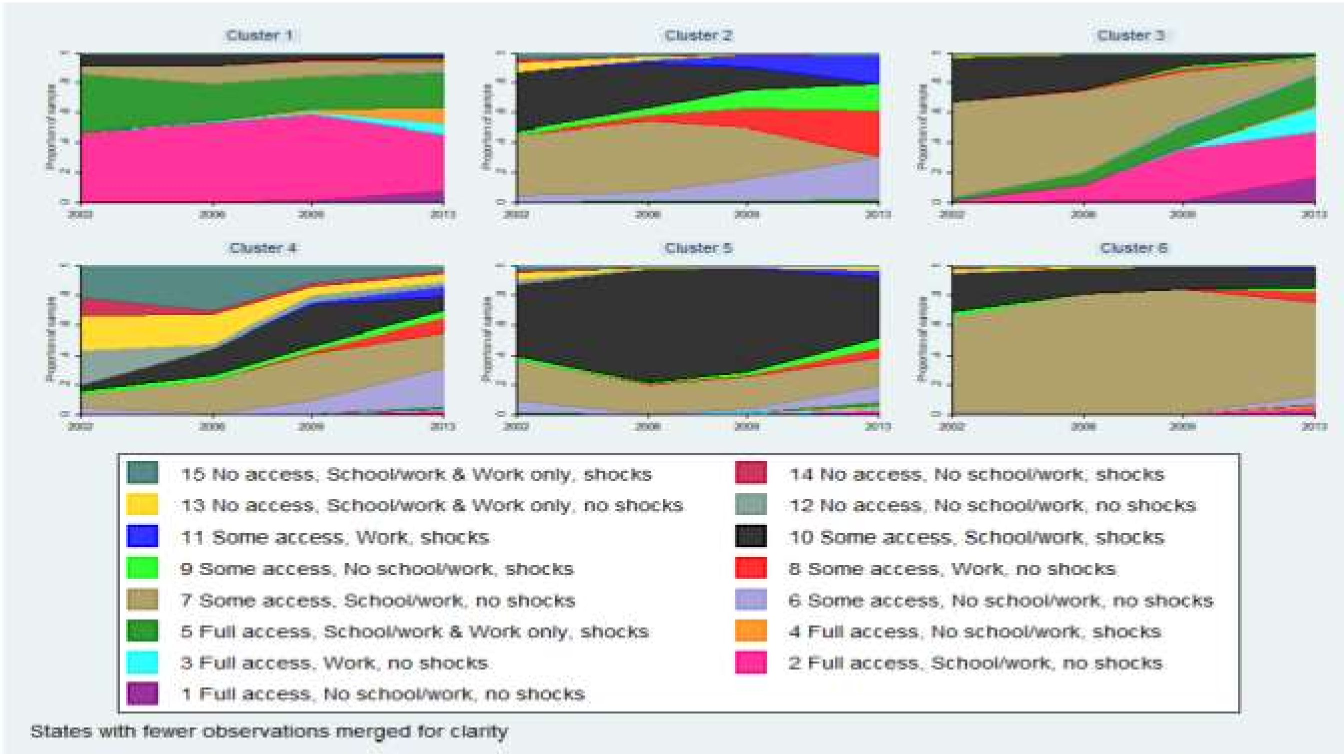


Figure 2: *Time-state Distribution Plots by Cluster*



DATA APPENDIX

Table A1. Definition of variables used in the regression analysis

Variable	Definition	Sample mean
CLUSTERj	Cluster grouping, j = 1,6.	3.71
LAST_ROUND	Last round of the survey (round 4)	0.25
LAST_ROUND*Cluster 2	Cluster 2 members in last round of survey.	0.06
LAST_ROUND*Cluster 3	Cluster 3 members in last round of survey.	0.02
LAST_ROUND*Cluster 4	Cluster 4 members in last round of survey.	0.03
LAST_ROUND*Cluster 5	Cluster 5 members in last round of survey.	0.04
LAST_ROUND*Cluster 6	Cluster 6 members in last round of survey.	0.07
Health	Subjective health of the child measured on a scale of 1-3 (1 = very poor/poor; 2 = average; 3 = good/very good)	2.38
Wellbeing	Self-rated wellbeing/life-satisfaction of the child measured on a scale of 1-9. 1 represents the worse possible life, 9 the best possible life lived. Constructed from eight questions related to different aspects of the child's life; opportunities for education and work, access to support from neighbours/friends, access to formal support and health services, quality of housing and living environment, food availability	4.94
Female	Female gender = 1; male gender = 0.	0.49
Age	Age of the child in months.	162.85
Education of household head	Educational attainment of the household head on a scale 1-19; no education = 0; masters/doctoral degree = 19	7.25
Age of household head	Age of household head measured in years.	44.21
Female household head	Female household head = 1; Male household head = 0.	0.19
Rural location	Residential location of the child. Rural = 1, Urban = 0.	0.65

Notes:

The sample means confirm that cluster 6 is the largest cluster and indicate that reported mean health was average to good/very good and similarly life-satisfaction was rated just above mid-way on the self-rated scale. Males and females are almost equally represented in the sample with marginal under-representation of females (see also below on survey and sample attrition). The means age of the sample is 13.6 years. Unsurprisingly, female headed household are in the minority. The Young Lives survey aimed to over sample children in poorer regions and this is reflected in the over-representation of rural households in the sample and the relatively low mean education of household heads (7 on the scale records Grade 7/Year 1 Secondary).

Analysis of attrition rates in the sample indicates that attriting children were mostly from households were heads had lower educational attainment (indicated by grade completion). As less educated heads are more often in poorer households this appears to be consistent with the Young Lives claim that attrition was linked to concerns that the study had not generated any tangible benefits for poorer households. It may also imply that the sample does not over-represent poorer people to the same extent as the overall Young Lives survey. We also find that despite boys having a higher attrition rate than girls, the proportion of girls attriting increased throughout the survey collection round. For instance, in round 2, 41.3% of the total number of children that attrited were girls attrited. This increased to 47.6% and 49.7% in rounds 3 and 4 respectively. This could be linked to marriage (which Young Lives say is a factor in attrition) as girls are more likely to leave their parent's homes when they marry (Levine and Kevane, 2003). However, 49.31% of the sample are female compared with 49.27% of all older cohort children surveyed in round 1, suggesting that differences in the attrition rates of girls and boys have not unbalanced the sample.

Table A2. Distribution of main states for the whole sample and by country

No. of observations (%)	Whole sample n=12,256 (100)	Ethiopia n=3,600 (29.37)	India n=3,768 (30.74)	Peru n=1,444 (11.78)	Vietnam n=3,444 (28.10)
Living conditions/Access to amenities					
Full access	2133 (100)	71 (3.33)	952 (44.63)	749 (35.11)	361 (16.92)
Only some access	9250 (100)	2919 (31.56)	2761 (29.85)	647 (6.99)	2923 (31.60)
No access	873 (100)	610 (69.87)	55 (6.30)	48 (5.50)	160 (18.33)
Economic status: School and/or paid work					
No school or paid work	1593 (100)	648 (40.68)	446 (28.00)	118 (7.41)	381 (23.92)
School with/without paid work	9,723 (100)	2786 (28.65)	2955 (30.39)	1224 (12.59)	2758 (28.37)
Paid work only	940 (100)	166 (17.66)	367 (39.04)	102 (10.85)	305 (32.45)
Shocks					
No shock	7496 (100)	1825 (24.35)	2579 (34.41)	798 (10.65)	2294 (30.60)
Family shock only	3877 (100)	1370 (35.34)	1086 (28.01)	455 (11.74)	966 (24.92)
Economic shock only	477 (100)	170 (35.64)	53 (11.11)	139 (29.14)	115 (24.11)
Economic and family shock	406 (100)	235 (57.88)	50 (12.32)	52 (12.81)	69 (17.00

Supplementary Materials

Supplementary Table S1. Full distribution of states

Main activity in recorded year	Code	Frequency	Percent
1 Full access, No school/work, No shocks	AMS	121	1.05
2 Full access, School/work, No shocks	ANS	1,158	11.09
3 Full access, Work, No shock	AOS	96	0.9
4 Full access, No school/work Family shock	AMT	64	0.54
5 Full access, School/work, Family shock	ANT	467	4.25
6 Full access, Work, Family shock	AOT	46	0.47
7 Full access, No school/work, Economic shock +/- family shock	AMU	21	0.18
8 Full access, School/work, Economic shock +/- family shock	ANU	148	1.23
9 Full access, Work, Economic shock +/- family shock	AOU	12	0.09
10 Some access, No school/work, No shocks	BMS	694	5.39
11 Some access, School/work, No shocks	BNS	4,476	36.16
12 Some access, Work, No shock	BOS	488	3.84
13 Some access, No school/work, Family shocks	BMT	369	2.91
14 Some access, School/work, Family shock	BNT	2,334	18.3
15 Some access, Work, Family shock	BOT	239	1.83
16 Some access, No school/work, Economic shock +/- family shock	BMU	62	0.46
17 Some access, School/work, Economic shock +/- family shock	BNU	554	4.2
18 Some access, Work, Family &/or economic shock	BOU	34	0.3
19 No access, No school/work, No shocks	CMS	155	1.17
20 No access, School/work, No shocks	CNS	294	2.44
21 No access, Work, No shock	COS	14	0.1
22 No access, No school/work, Family shock	CMT	90	0.67
23 No access, School/work, Family shock	CNT	259	1.96
24 No access, Work, Family shock	COT	9	0.07
25 No access, No school/work, Economic shock +/- family shock	CMU	17	0.13
26 No access, School/work, Economic shock +/- family shock	CNU	33	0.26
27 No access, Work, Economic shock +/- family shock	COU	2	0.01
Total		12,256	

Notes: Full(A)/Some(B)/No(C) access: Home has access to all/some/none of electricity, own toilet, piped drinking water, adequate fuels for cooking. No School/work (M); neither in school nor work. School/work (N); in school and not working or also working (states combined as very few observations in which child is in school and working (n=363)). Work (O); working only.

No shock (S); no family or economic shock suffered. Family shock (T); suffered family shock (divorce, separation, family death or illness) but not any economic shock; Economic shock +/- family shock (U); suffered economic shock (loss of employment or source of income or family enterprise) either with or without also suffering a family shock (states combined as relatively few observations in which an economic shock is suffered without a family shock (n=501))

Supplementary Table S2. Interpretation of DiD interaction effects in ordered logit HEALTH and WELLBEING estimates: Marginal effects at baseline and last round and difference-in-differences.

Cluster	HEALTH Baseline	HEALTH Last Round	Difference-in-differences	WELLBEING Baseline	WELLBEING Last round	Difference-in-differences
2	-0.0646	-0.0584	+0.0064	-0.0629	-0.0253	+0.041*
3	0.0131	-0.0673	-0.0804*	-0.0112	0.0004	+0.0211**
4	-0.0895	0.0150	+0.0823***	-0.0574	-0.0252	+0.0826**
5	-0.0520	-0.0419	+0.0101	-0.0457	-0.0229	+0.0686**
6	-0.0088	-0.0266	-0.0178	-0.0202	-0.0120	+0.0322**

Notes: Marginal/discrete effects of clusters calculates with as observed values of all the other covariates (average marginal effects) for the probability of reporting (1) good/very good health and (2) the mean level of wellbeing (5) relative to cluster 1. The difference-in-differences are the discrete changes between the marginal effects of each cluster at the baseline and last round A positive (negative) difference-in difference represents a more (less) positive marginal effect in the last round i.e. a narrowing (widening) difference/gap. A positive marginal effect in the last round indicates that the health/wellbeing gap has closed.

Supplementary Table S3. Robustness test results for alternative health measure

		(1a) OLS	(1b) O. Logit
Independent variable	Dependent variable	HEALTH2 (β)	HEALTH2 (e^{β})
LAST_ROUND		0.00341	1.100
		(0.128)	(0.380)
Cluster 2 *Early transition to adult states*		-0.171***	0.602***
		(0.0524)	(0.0857)
Cluster 3 *Transitioning to better-off*		-0.00353	1.000
		(0.0593)	(0.160)
Cluster 4 *Poor-to-average & some instability*		-0.176***	0.597***
		(0.0682)	(0.112)
Cluster 5 *Average but unstable*		-0.167***	0.630***
		(0.0545)	(0.0944)
Cluster 6 *Average & stable*		-0.0821	0.798*
		(0.0499)	(0.108)
LAST_ROUND*Cluster 2 *Early transition to adult states*		0.00764	1.061
		(0.0643)	(0.184)
LAST_ROUND*Cluster 3 *Transitioning to better-off*		-0.102	0.726
		(0.0821)	(0.160)
LAST_ROUND*Cluster 4 *Poor-to-average & some instability*		0.261***	2.041***

	(0.0809)	(0.453)
LAST_ROUND*Cluster 5 *Average but unstable*	0.126*	1.358*
	(0.0680)	(0.252)
LAST_ROUND*Cluster 6 *Average & stable*	0.0115	1.003
	(0.0616)	(0.166)
Female	-0.183***	0.620***
	(0.0199)	(0.0338)
Age (in months)	-0.0346	0.895
	(0.0295)	(0.0717)
Education of household head	0.00568**	1.014*
	(0.00268)	(0.00737)
Female household head	-0.0830***	0.796***
	(0.0264)	(0.0579)
Age of household head	-0.000681	0.998
	(0.000990)	(0.00271)
Rural location	-0.00164	1.029
	(0.0290)	(0.0826)
India	-0.234***	0.415***
	(0.0333)	(0.0391)
Peru	-0.549***	0.166***
	(0.0387)	(0.0183)
Vietnam	-0.768***	0.0895***
	(0.0319)	(0.00859)
Constant/Cutpoint 1	4.937***	-9.296
	(0.449)	(1.237)
Observations	5,149	5,149
R^2 / Pseudo R^2	0.179	0.099
F/LR Chi^2	55.75***	1186.58***

Notes: HEALTH2 records only self-rated child health reported in survey rounds 3-4.
Reported figures are coefficients (β) for OLS and odds ratios (e^{β}) for ordered logit.
Only cutpoint 1 value reported for ordered logit.
Standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.