Young Lives' Constructed files

The constructed files are combined sub-sets of selected variables from Round 1, 2 and 3 of the Young Lives survey. The files contain about 200 original and constructed variables, most of them comparable across the three rounds, presented in a panel format and classified in four broad groups: panel information, general characteristics, household characteristics, and child characteristics. This document is organised around the same groups.

1. Panel information

Data – Three dummy variables showing whether the child has been found in each round, and an additional dummy signalling whether child was present in all rounds.

inr1	Child is present in Round 1
inr2	Child is present in Round 2
inr3	Child is present in Round 3
panel123	Child is present in all rounds

2. General characteristics

Data –Date of interview, and other basic geographical identifiers (area of residence, region of residence, sentinel site id, and community id).

dint	Date of interview
commid	Community id
clustid	Sentinel site id
typesite	Area of residence (urban/rural)
region	Region of residence

3. Household characteristics

3.1.Household wealth

Data – Wealth index and constituent sub-indexes (housing quality, access to services, and consumer durables).

The indexes were estimated consistently across rounds. For this purpose, only variables common to the three rounds were included. The wealth index is composed of three sub-indexes: (i) housing quality index (hq), (ii) access to services index (sv), and (iii) ownership of consumer durables (cd); all of which have equal weights in the estimation of the wealth index. Therefore, the wealth index of household i will be defined as:

$$wi_i = \frac{hq_i + sv_i + cd_i}{3}$$

Housing Quality Index - is the simple average of the following indicators:

- Crowding (scaled sleeping rooms per person)
- Main material of walls – dummy variable that takes the value of 1 if main material of walls satisfied basic norms of quality.
- Main material of roof
- Main material of floor

Access to services – simple average of the following indicators:

- Access to electricity
- Access to safe drinking water
- Access to sanitation
- Access to adequate fuels for cooking

Consumer durables – simple average of a set of dummy variables which take the value of 1 if a household member owns at least one of each consumer durable. To ensure comparability across rounds, only those consumer durables common to all three rounds were included. The following tables show the lists of common consumer durables used in each country:

Ethiopia – 10 common items across rounds

Radio	Landline phone
Television	Mobile phone
Bicycle	Table and chair
Motorbike	Sofa
Automobile	Bedstead

India – 9 common items across rounds

Radio	Landline phone
Television	Mobile phone
Bicycle	Refrigerator
Motorbike	Fan
Automobile	

Peru - 12 common items across rounds

Radio	Landline phone
Record player	Mobile phone
Television	Refrigerator
Bicycle	Iron
Motorbike	Blender
Automobile	Stove

Vietnam – 9 common items across rounds

Radio	Landline phone
Television	Mobile phone
Bicycle	Refrigerator
Motorbike	Fan
Automobile	

3.2.Consumption aggregates

Data – Total per capita expenditure, per capita food consumption, and per capital non-food expenditure, all in both nominal and real terms.

The construction of the consumption aggregates involved adding together a number of items grouped in two main classes: (i) food items, and (ii) non-food items. It should be noted that whilst a core set of items is similar in all four countries, other food and non-food items that are specific to each country were added in the design of the questionnaire.

- Food consumption.
  Aggregation of all food items consumed in the last 2 weeks from different sources: (i) food purchased; (ii) food home-produced (from own harvest) or from stock; (iii) food items received as gifts or transfers; (iv) food received from employers as payment in-kind for services rendered. In the case of Peru, there is an additional question related to all food that was left over. Therefore, this amount is subtracted from the final aggregate.
  Each source is converted to monthly terms (by multiplying it by 2 because the recall period is 2 weeks) and finally aggregated.

- Non-food consumption
  Aggregation of all non-food items, which are classed together in 4 big groups: (i) expenditure on education; (ii) expenditure on health; (iii) expenditure on clothing and footwear; and expenditure on other non-food items. The selection of the items was based on a comparability criteria (i.e. all those items that were included consistently in all rounds). Since this information was collected for different reference periods, they all were converted to months before aggregating them.

  (i) Expenditure on education. Includes all money spent on school uniform for boys and girls, payments for tuition, fees or donations to school, books and stationary, and transport to school.
  (ii) Expenditure on health. Includes all money spent on medical consultations and treatment and other medical expenses.
  (iii) Expenditure on clothing and footwear. Includes all money spent on clothing and footwear for adults and children.
  (iv) Expenditure on other non-food items. Includes all money spent on other non-food items such as: rents, dwelling and vehicle maintenance, water supply, electricity rates, telephone and mobile phone rates, fees and paperwork, legal

advice, bribes, one-off family events, festivals and celebrations, personal care items, entertainment, presents for children, and jewellery (although this item was excluded from the consumption aggregate of Peru and Vietnam because it caused too much distortion).

Once food and non-food items are aggregated, they are adjusted for cost of living differences and for household composition. Prices to deflate nominal consumption aggregates are collected from external sources for all countries, except for India, where the information comes from community questionnaire, as it seemed more appropriate for our research purposes. This information is now archived together with the main data in order to be able to reproduce the same consumption aggregates if needed.

In all countries, except Ethiopia, the results are adjusted by the household size (i.e. all members that live permanently in the household) and, thus, are reported in per capita terms. In Ethiopia, the results are reported in 'per adult' terms. This means that real expenditure is divided by the household size adjusted for adult equivalence. Dercon (1995) proposed the following equivalences based on nutritional (caloric) requirements of different ages and men and women.

Years of age	Men	Women
0-1	0.33	0.33
1-2	0.46	0.46
2-3	0.54	0.54
3-5	0.62	0.62
5-7	0.74	0.70
7-10	0.84	0.72
10-12	0.88	0.78
12-14	0.96	0.84
14-16	1.06	0.86
16-18	1.14	0.86
18-30	1.04	0.80
30-60	1.00	0.82
60 plus	0.84	0.74

3.3. Livestock ownership

Data – Number of animals owned by the household.

In order to have comparability across rounds, animals were aggregated in four big groups: milk animals (*animilk*), draught animals (*anidrau*), small ruminants (*anirumi*), and animals that are specific to each country (*anispec*). Additionally, for Round 2 and Round 3, the number of livestock is reported for those animals that were asked consistently in both rounds for each country.

3.4. Land ownership

Data – access to land (owned, borrowed, rented, etc.) in the last year (*accland*), total area of land the household has had access to (in hectares) in the last year (*totland*), total area of land owned by the household (in hectares) in the last year (*ownland*).

In India, land area was collected consistently in acres. Therefore, for the panel datasets, the information was transformed to hectares using the conversion factor 1 acre= 0.4047 hectares. In Vietnam, the unit in which land extension was collected was squared metres. For this reason, the conversion factor used was 0.0001 (to hectares). Note that in Vietnam information about total area owned refers to having a long-term use certificate for the land.

In Ethiopia, whilst in Round 1 this information was recorded in squared metres, in Round 2 and Round 3, it was recollected using (most commonly used) local units. Therefore, all the information was harmonized by converting it to hectares. This was done by using the following conversion factors:

1 unit of	In hectares
Gasha	40.00
Timad, Gemed, Kert, Kedema	0.25
Massa	0.67
Kufaro	0.03
Zhir	0.0003
Boy	0.005
Tinto	0.06
Ermija, squared metres	0.0001
Dearo	0.045
Gezem	0.1666
Kend	0.000025
Fer	0.25

Similarly in Vietnam, Round 2 and Round 3 collected information on land using local units. The conversion factors are listed in the following table:

3.5. Characteristics of the household head

Data – basic characteristics of the household head such as the ID in the roster, age, sex, level of education (i.e. highest grade attained).

Because it is expected that household heads maintain their level of education across rounds, unless they were enrolled in formal schooling and had attained new grades as a consequence, the variable for the level of education has been harmonized for those households that have the same household heads throughout the three waves of interview or in at least two of them. In this way, we overcome reporting inconsistencies and we end up with only one level of education for a singular household head.

In the case of Peru, round 2 information on the highest level of education was considered for all rounds.

3.6.Characteristics of the primary caregiver

Data – basic characteristics of the primary caregiver such as the ID in the roster, age, sex, literacy, level of education (i.e. highest grade attained), relationship to household head, relationship to index child, and an additional variable related to caregiver's subjective well-being (in terms of the nine steps of a life-satisfaction ladder).

The variable of education was harmonised in the same way as for the household head (see 3.5).

careid	Caregiver's ID in roster
careage	Caregiver's age
caresex	Caregiver's sex
carehead	Caregiver's relationship to head of household
carerel	Caregiver's relationship to child
carelit	Caregiver is literate
caredu	Caregiver's level of education - harmonized var
ladder	Caregiver's subjective well-being (ladder)

Caregiver relationship to the index child (*carerel*) has been harmonized across rounds. It has 6 options: (1) biological parent, (2) partner of biological partner, (3) Grandparents, (4) Uncles/Aunts, (5) Siblings (which includes half-siblings), and (6) Other (including all other household or non-household members).

Caregiver id (*careid*) is reported for all rounds and cohorts, except for Vietnam that did not collect information about the main caregiver for the OC in R3.

3.7.Household composition

Data – household size (including the index child), and number of household members by sex and age groups (excluding index child).

hhsize	Household size
male05	Number of males aged 0-5
male612	Number of males aged 6-12
male1317	Number of males aged 13-17
male1860	Number of males aged 18-60
male61	Number of males aged 61+
female05	Number of females aged 0-5
female612	Number of females aged 6-12
female1317	Number of females aged 13-17
female1860	Number of females aged 18-60
female61	Number of females aged 61+

3.8.Shocks

Data – occurrence of events that have affected negatively the economic situation of the household.

All shock-related variables are binary, being 1: shock was reported, 0: shock was not reported. Note that answers are based on perceptions, this is, they do not show whether a negative event has occurred or not, rather they show whether the respondent considers the event has affected the welfare of the household negatively.

4. Child characteristics

4.1.General characteristics

Data – sex, age (in completed months), first language, ethnicity, and religion.

Given that sex, ethnicity, and religion are time-invariant variables, they were taken from Round 1.

Age in months is estimated by taking the age of the child in days (date of Interview-date of birth) and dividing this number by 365/12 (number of days per month). The final number is rounded up to one decimal point. In order to preserve anonymity, dates of birth cannot be publicly archived, therefore external users will not be able to estimate this variable, but details on how it was estimated can be found in the do-files.

Although child's first language is a time-invariant variable, it was taken from Round 2 because the information was not collected in Round 1. Therefore, most missing values in this variable are explained by attrition.

4.2.Child health and nutrition

Data – Anthropometric information, weight at birth, antenatal care, constructed measures for malnutrition, and self-reported health (i.e. in relation to other children of same age, serious injuries, and long-term health problems).

Anthropometric information includes health and weight, and z-scores for weight-for-height, height-for-age, and bmi-for age. The z-scores were estimated using WHO references tables and software (available for download at: http://www.who.int/childgrowth/en/). These measures were estimated using the age of children in days. To keep anonymity, the latter information cannot be publicly archived; therefore, the results we provide cannot be reproduced exactly. Age of child in months rounded up to one decimal, however, provides very close estimators.

Weight at birth and antenatal care information was only asked to the Younger Cohort in Round 1.

Malnutrition estimators were constructed on the basis of the z-scores. The six estimators that are included in the panel files are: stunting, severe stunting, thinness, severe thinness, underweight, and severe underweight. The following table provides the definition for each:

Stunting	<-2 SD of height-for-age z-score
Severe stunting	<-3 SD of height-for-age z-score
Underweight	<-2 SD of weight-for-age z-score
Severe underweight	<-3 SD of weight-for-age z-score
Thinness	<-2 SD of bmi-for-age z-score
Severe thinness	<-3 SD of bmi-for-age z-score

Finally, three variables of self-reported health are included:

(1) Child's health relative to other children of the same age (*chhrel*).
Although this information exists for all rounds and cohorts, the respondents and the scales have changed.

Rounds/Cohort	Respondent	Scales
R1-YC	Caregiver	Same, better,worse
R1-OC	Caregiver	
R2-YC	Caregiver	
R2-OC	Caregiver, child	
R3-YC	Caregiver	Much worse, worse, same, better, much better
R3-OC	Child	

In order to have panel information a new variable was created using caregiver's responses for R1 and R2, and R3-YC and the child's response for R3-OC. Additionally, the scales of R3 were reduced to three –same, better, worse- by aggregating much worse and worse, and much better and better, as worse and better, respectively. Users are asked to keep this in mind when using this variable.
In India, this question was omitted in R3, therefore the panel datasets only show information for R1 and R2.

(2) Serious illnesses and injuries (*chhilin*).
This information was collected consistently in Round 1 and Round 2; however, in Round 3, the question was somewhat different, focusing only on serious injuries (instead of asking for serious injuries and illnesses as in R1 and R2). For this reason, the panel dataset will only include the first two rounds of data.

The specific question that was asked was: 'in the last [##] years, has the child had a serious illnesses or injuries, when you REALLY thought he/she might die?'. The constructed variable is dichotomic, with 1 meaning that child has had a very serious illness or injury, and 0; otherwise.

(3) Long-term health problems (*chhprob*).
Similarly, information related to this variable was collected with some variation across rounds. In Rounds 1 and 2, the question was straightforward: 'Does child have any long-term health problem that affects how he/she attends to school or work?' However, in Round 3, this question was omitted and instead a list of long-term health problems was asked, in two steps: first, asking whether the child suffers from any of the listed long-term health problems, and second, asking if the problem affects his/her capacity to study.
For this reason, the panel variable corresponds to Round1 and Round 2 only; a dummy that takes the value of 1 if child has long term health problem, and 0 otherwise.
Additionally, for Ethiopia, a set of dummy variables is added for specific long-term health problems listed in Round 3. These are: poor vision, hearing problems, frequent headaches, and respiratory problems (*prvision, prhear, prhead, and prrest,* respectively).

4.3.Child education/ formal skills

Data – enrolment, problems with reading and writing, literacy, school type, highest level attained, travel time to school, raw and standardised scores of PPVT, and number of correct answers in Ravens Test.

Enrolment (*enrol*) is constructed based on information on whether the child was attending formal school at the time of the interview. For Round 1, this information was collected for the older cohort only (the younger cohort was too young to be enrolled in school). In Round 2, when the younger cohort is about 5 years old, enrolment includes pre-school enrolment. Following this, if the child was 'still attending pre-school,' at the time of the interview, he/she will be counted as enrolled in school.

Problems with reading and writing are based on the reading and writing items of the achievement and development instruments. For each of the items, a test was administered to the child. After the finalization of the tests, fieldworkers are asked to fill in some answers based on what/how the child responded, as follows:

Reading item
01= Can't read anything
02= Reads letters
03= Reads words
04= Reads sentences
Writing item
01= No (can't write anything)
02= Writes with difficulties or error
03= Yes, without difficulties or errors

Based on this information, problems with reading (*readprob*) accounts for all those cases where children are not able to read sentences (!=04). Problems with writing (*writeprob*), in turn, will be defined as not being able to write without difficulties or errors (!=03).

In close relation to the two previous variables, literacy (*literate*) is defined as the ability to read and write without problems.

The school type variable (*schtype*) is a scaled-down version of the original that includes a range of options (private, public, community, government funded, community, NGO-Charity, etc). The variable in the panel file only shows the options public, private, and other, for all countries. For Round 2, the type of pre-school is included for all those children enrolled in pre-school at the time of the interview. In Round 3, the question related to school type was not asked. For the Older Cohort, however, yearly school information is included in the school history. For the purposes of the panel dataset, the type of school that the child was attending the year of the interview (2009) is included. The only exception is Vietnam where school history was not collected in R3.

Ethiopia R1- question only referred to public and private.

The grade of school (*chgrade*) refers to the highest grade completed by the child at the time of the interview. For children who never attended school, the variable will have the value of 0.

In terms of the Ravens Test (*ravens*), administered to the Older Cohort in round 1 only, information was not collected uniformly across countries; whilst in Peru and India the test was administered to the whole cohort, in Vietnam it was administered only to a sample of 200 children living in urban areas. In Ethiopia, in turn, the test was not administered at all.

4.4.Time allocation

Data – number of hours that the child spends performing different activities on a typical day.

Time-allocation information was not collected in Round 1. For Round 2 and Round 3 the information was collected at the household level -to all household members age 5-17- and at the child level (for the older cohort only). To keep consistency, in the panel files, time-use

will consider the household level information only (i.e. that answered by caregiver or other well-informed adult). Since the information was collected for household members age 5-17, missing values in Round 2, younger cohort, correspond to children younger than 5 at the time of the interview.

The activities around which time-use information is collected are:

- Sleeping - *hsleep*
- Caring for others in the household - *hcare*
- Running household chores – *hchore*
- Working on household tasks (farming, herding, etc.) – *htask*
- Working outside household on paid activities – *hwork*
- At school – *hschool*
- Studying outside school (doing homework, extra classes, learning languages, etc.) – *hstudy*
- Playing, leisure time (including eating, showering, etc.) – *hplay*

4.5.Subjective and psychosocial well-being

Data – Child's self-assessment of personal wellbeing in terms of a nine-step ladder.

The information displays the step number in which the child situates his/her personal well-being at the time of the interview, where 9 represents the 'best possible life' and 1 'the worst possible life'. This information was collected in Round 2, for the older cohort only, and in Round 3 for both cohorts.

4.6.Parental characteristics

Data – for both biological mother and father the information in the panel file is:

- ID in roster
- age
- literacy - taken from Round 2
- highest grade attained– harmonized across rounds
- presence in household (at the time of the interview)