Young Lives

INSIGHTS REPORTS

May 2021

Leveraging the Power of Longitudinal Data: Insights on Data Harmonisation and Linkage from Young Lives

Jo Boyden and Deborah Walnicki

Introduction

This report focuses on data harmonisation and data linkage. It makes the case that these procedures have the potential to help address some of the most pressing data constraints currently facing researchers and policymakers in low- and middle-income countries (LMICs). Despite considerable investment and notable improvements in recent decades, data limitations in LMICs continue to undermine the advancement of scientific knowledge and the design of effective policies, services and programmes. There are multiple data challenges. For example, commenting on the status of education data, a recent Center for Global Development blog cites small samples, limited accuracy and irrelevance of data, together with their disjointed nature, as well as access restrictions and weak capacity for analysis (Rossiter 2020). Additionally, most of the instruments, measures and indicators deployed in social and medical research internationally have been designed in the Global North, which can hamper efforts to identify measures that reliably, validly, and fairly assess the traits of interest across LMIC contexts where very different material circumstances and sociocultural norms apply (Dawes 2020).

Reflections on

- how data harmonisation and linkage can address data constraints in low- and middle-income countries and increase the possibility that policy relevant questions are answered.
- how data harmonisation can increase the sample size and statistical power of a study, as well as increase the generalisability of findings.
- how the process of data linkage can allow researchers to access expanded variables and increase opportunities for innovation.
- the cost effectiveness and limitations of both data harmonisation and linkage processes.

There is therefore a huge need to improve both the quality and the consistency, as well as the comparability and integration, of LMIC data. While this need applies to all forms of data, longitudinal research confronts specific challenges, due not only to the technical and logistical complexity, but also the cost and skills requirements, of panel data collection and analysis, and the mismatch between the time involved in generating robust findings and policy stakeholders' wishes for instant evidence. Additionally, there are perceived limitations to the policy applicability of much existing longitudinal evidence, especially its relevance across contexts, populations and treatments, not just because of the small sample sizes and lack of national representativeness, but also the risk of cumulative bias due to selective participant attrition (Curran et al. 2008). Yet longitudinal evidence offers unique scientific and policy advantages in LMICs. In particular, by tracking the same individuals at regular intervals over time, longitudinal cohort research sheds light on the causal relationships in human development and well-being, tracing the factors and mechanisms that shape individuals' trajectories and outcomes. Longitudinal cohort data can also illuminate the impact of policies and interventions.

Data harmonisation and data linking can be an important asset to longitudinal research in LMICs, extending the reach and enhancing the impact of longitudinal evidence at limited additional cost. Focusing mainly on cohort research, this paper enumerates the aims and approaches of recent harmonisation and linking initiatives and underlines some of the opportunities and challenges involved. In particular, it highlights learning from Young Lives, a longitudinal study of 12,000 young people in Ethiopia, India,[1] Peru, and Vietnam. The paper concludes with a summary of the key advantages of and lessons learned from data pooling and data linking.

Definitions: data harmonisation and data linkage

Since different naming conventions are used to refer to these procedures, it is important to highlight the definitions used in the paper. *Data harmonisation,* or *data pooling,* is understood to entail the amalgamation of diverse types and sources of data so that they are sufficiently compatible and comparable to be integrated within a single dataset and analytical framework.[2] Data pooling extends a sample by adding new observations/subjects, thereby increasing statistical power and potentially also permitting greater heterogeneity in demographic representation. There are two broad approaches to data harmonisation. *Prospective harmonisation* is when the research questions, methodology and comparable outputs across methods of inference are, as far as possible, planned in advance, enabling analysis of comparable outputs to be undertaken directly, without requiring further harmonisation procedures. *Retrospective,* or *post hoc, harmonisation*, entails the pooling of data when a priori standardisation of different research instruments is not feasible or data collection has already been completed.

Data linkage is taken to involve connecting variables or records from a study sample with variables or records on that same sample that are held by external sources. This makes it possible to expand or validate the information available on existing observations/subjects, thereby opening up the possibility of new lines of research. It is often used to match survey data with administrative data that are routinely collected by government and other institutions, permitting monitoring of access to education, health, and social provision, evaluation of the impact, and identification of the features of services that have the greatest effect. Another common use is connecting survey data with nationally representative datasets.[3] Recently, the expansion of roads and public amenities has led to increased interest in linking survey data with open access geospatial data, while concerns about environmental destruction and climate change have resulted in growing demand to bring survey data together with meteorological, pollution and biodiversity data.

Increased power of data harmonisation

Opportunities created by data harmonisation in longitudinal research

- Harmonising data from different longitudinal studies increases the sample size and statistical power of data.
- Data harmonisation increases the possibility of discerning whether conceptually equivalent relationships can be detected across population groups, settings, times, treatments and outcomes.
- Increased generalisability and statistical power mean that harmonised data have enhanced impact. This can strengthen evidence-based policy recommendations.

There are well over 100 longitudinal randomised controlled trials, quasi experiments and panel and cohort studies in LMICs. However, compared to similar studies in high-income countries, most struggle with logistical and resource constraints and are confined to relatively small samples and limited data points. Several initiatives in recent years have aimed at overcoming these obstacles and promoting longitudinal research in LMICs through increased collaboration, joint fundraising, enhanced data

1 In the states of Telangana and Andhra Pradesh.

2 Data are pooled by marshalling the variables of interest into a common format that measures the same latent constructs. For the analysis to be considered robust, the data must be inferentially equivalent irrespective of the instrument or data collection method used.

3 The Demographic and Health Survey offers nationally representative data on health and nutrition in 90 different countries, and is one of the datasets most often used in linkage exercises in LMICs. See https://dhsprogram.com

discoverability and knowledge exchange, among other activities. The reconciling of instruments and measures and pooling of data across studies is intrinsic to this work. For example, the INDEPTH Health and Demographic Surveillance System (HDSS) network engages in capacity strengthening and collaborative multi-centre longitudinal health and demographic research across a range of transitioning settings.[4] It aims to understand and improve population health and development policy, practice, and progress and expand the underlying longitudinal tracking platform. The comparative advantage of INDEPTH centres on conducting longitudinal population-based tracking, comparing data across different systems and cultures, and leveraging the collective value of data across sites. Similarly, established by UNICEF's Office of Research, GLORI is another network of longitudinal researchers that aims to improve the quality, utility, accessibility, consistency and comparability of longitudinal data on children and adolescents as a step towards improved policy and practice (UNICEF n.d.). Both INDEPTH and GLORI have two wider goals, which are to enhance the contribution of longitudinal evidence to the Sustainable Development Goals (SDGs) and strengthen institutional capacity for and local ownership of longitudinal research.[5] The Cohort and Longitudinal Studies Enhancement Resources (CLOSER) project also aims to increase data discoverability, identify potential areas of data harmonisation and linkage, and encourage collaborations between longitudinal studies in LMICs through CLOSER International (CLOSER 2018).[6] To this end, the team created the Low and Middle Income Longitudinal Population Study (LMIC LPS) Directory, which provides open-access information on more than 170 studies.[7]

In addition to these coordination efforts, there are several data pooling projects in LMICs that have the objective of creating multidimensional indices for employment across countries and time periods as a means of benchmarking standards in human health, development and well-being. Their central application is as a monitoring tool that can be used to inform and call decision-makers to account, such as around levels of expenditure on or the performance of services. For example, the global Multidimensional Poverty Index (MPI), which is based on repeated cross-sectional surveys in over 100 LMICs, aims to measure poverty across diverse dimensions by considering people's simultaneous and overlapping deprivations in health, education and living standards (Alkire, Kanagaratnam, and Suppa 2020). Created by the Oxford Poverty and Human Development Initiative (OPHI), MPI figures are updated annually and changes over time are computed using the global MPI, which relies on data from the Demographic and Health Surveys (DHS), Multiple Indicator Cluster Surveys (MICS) and some national surveys. The 2020 release included harmonised trends in MPI and related statistics for 80 countries (OPHI 2020).[8] Taken together with data disaggregated by subnational regions and social groups, the global MPI 2020 makes it possible to gauge who is poor and how they are poor, and to track progress on internationally agreed standards, such as SDG 1 to end poverty in all its forms and dimensions (Alkire et al. 2020).

To obtain comparable estimates across time points, each country's original MPI is adjusted to account for differences in the available data. By exactly aligning the indicator definitions across time, harmonisation of the MPI ensures that any differences observed are due to changes in the conditions of poverty. By introducing changes to reflect data availability across time, the harmonised MPIs are more tailored to each country and can have slightly different versions of the indicators, meaning that cross-country comparability is less straightforward (Alkire et al. 2020). Although the global MPI uses cross-sectional data due to availability, the harmonisation and trends analysis is equally possible using longitudinal data. In one study, researchers applied the Alkire-Foster measures to Young Lives data to understand how the children experienced multidimensional poverty (Apablaza and Yalonetzky 2013).

In another example, Weber et al. (2019) sought to enhance research globally on the development of children under 3 years old by creating a validated metric that is comparable across cultures and contexts. The team harmonised data for some 36,000 children retrospectively from 16 longitudinal cohort studies in 11 low-, middle- and high-income countries, and from this generated the Developmental Score (D-score), which represents a universal latent construct of early childhood development. Pooled item-level developmental assessment data for children age 0–48 months were derived from 12 internationally recognised and commonly used instruments that cover a broad range of longitudinal child development outcomes, including receptive vocabulary and reasoning. The data were combined into a single database using a set of linking items that performed equivalently across countries and cohorts, resulting in a matrix from which the D-score was constructed and validated (Weber et al. 2019).[9]

4 See http://www.indepth-network.org. The UK's Medical Research Council Lifecourse Epidemiology Unit also supports and collaborates with several longitudinal research groups in LMICs, including COHORTS, which encompasses five LMIC birth cohorts.

5 GLORI is creating an inventory of resources, harmonising tools where feasible and documenting and sharing best practice.

6 See https://www.closer.ac.uk/home

7 The directory is searchable by geographic location and/or topic and provides a summary of the details of each of the LMIC LPS included, as well as links to individual study websites, where further information is available. See https://www.ifs.org.uk/tools_and_resources/longitudinal?page=1

8 National MPIs can rely on existing surveys, in which case the choice of indicators is constrained to what is available in those data. In some cases, countries can decide to adapt an existing or create a new questionnaire that includes questions related to selected MPI indicators.

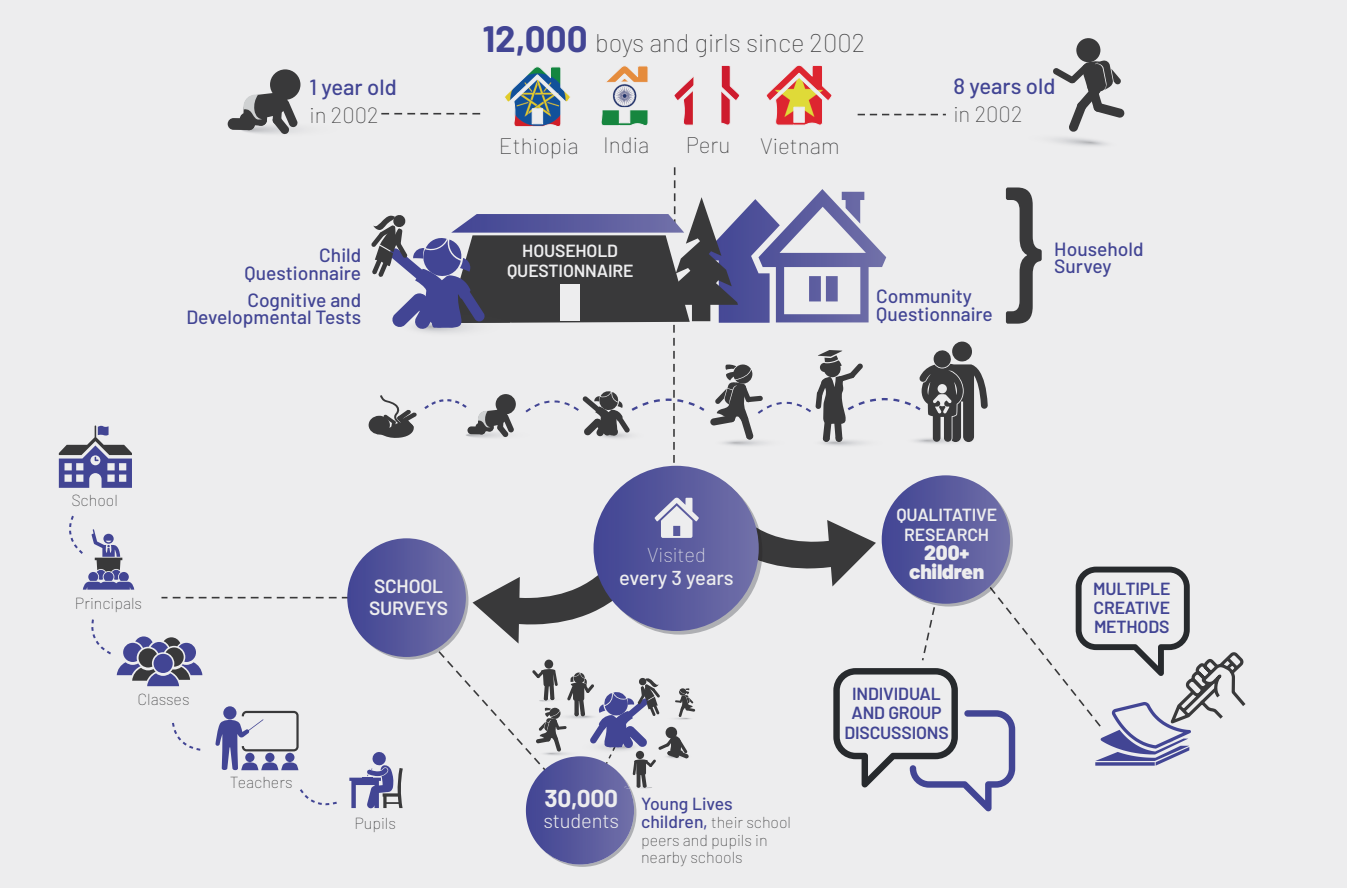
9 The authors describe how item mapping produced 95 'equate groups' of same-skill items across 12 different assessment instruments. A statistical model was built using the Rasch model with item difficulties constrained to be equal in a subset of equate groups, linking instruments to a common scale, the D-score, a continuous metric with interval-scale properties. D-score-for-age z-scores (DAZ) were evaluated for discriminant, concurrent and predictive validity to outcomes in middle childhood to adolescence.

For the Consortium of Health-Orientated Research in Transitioning Societies (COHORTS) study (Richter et al. 2012), the aim of pooling data retrospectively across different samples, country contexts and epochs was to be able to analyse the long-term health and human capital outcomes of maternal and child undernutrition (Victora et al. 2008; Adair et al. 2013). Data were pooled from what were, at the time, the five largest prospective birth cohort studies in LMICs, all of which had an initial sample size of 2,000 or more new-born children. The additional conditions of inclusion were that the studies must have started recruitment during gestation or at delivery, had at least 15 years of follow-up and a high prevalence of maternal and child undernutrition (Ader et al. 2013). The data covered maternal height, birthweight, intrauterine growth restriction, and weight, height, and body-mass index at 2 years old (according to WHO growth standards), and outcome variables were maternal height, schooling, income or assets, offspring birthweight, body-mass index, glucose concentrations, and blood pressure. Research questions, constructs and measures were aligned and data harmonised post hoc to ensure the consistency of findings across studies (Richter et al. 2012).

The integrated COHORTS study embodies many strengths over and above those of the original studies, and the pooled data were used to great effect in a series of highly influential journal articles (e.g. Fall et al. 2016; Fall et al. 2015; Ader et al. 2013). The enlarged sample ensured increased statistical power, offered greater heterogeneity in participant demographics and also made it possible to confirm individual site findings through verification with observations from other sites, conduct important checks of measurement invariance over development and across sub-groups, and test a much broader range of theoretically driven research hypotheses. Additional funding obtained recently has increased the temporal analysis considerably, with longitudinal follow-ups of the sample that extend well into participants' mid-adult years.

Similarly, post hoc data harmonisation was employed by Das, Singh and Yi Chang (2020) as part of the Research on Improving Systems of Education (RISE) Programme in order to research whether the attainment of human capital alone was enough to reduce inequality, increase mobility and ultimately address poverty.[10] Harmonising data from the longitudinal study Learning and Education Achievement in Punjab Schools Project (LEAPS) in Pakistan with data from the Young Lives four-country sample allowed the researchers to interrogate whether higher test scores in mathematics and receptive vocabulary at the end of primary school increase future higher education enrolment and attendance across five LMICs. The two studies contain measures of socio-economic background in childhood – specifically, ownership of household consumer durables and access to services – and have sufficient follow-up to

Figure 1. *Young Lives research components*



10 See https://riseprogramme.org

capture final schooling levels, the two dimensions being linked through a comparable metric over time.[11] The enlarged sample arising from data harmonised across five countries served to strengthen the study's policy recommendations. The researchers found that focusing solely on improving poor children's test scores did not equalise their access to college, their recommendation being that more needs to be done to close the gap in college enrolment for lower-income youth, including targeted access programmes for higher education (Das, Singh and Yi Chang 2020).

As a longitudinal study created to shed light on the predictors and outcomes of child poverty in four LMICs, prospective harmonisation of data across time and country context is an essential feature of the Young Lives research design. Figure 1 illustrates the key components of Young Lives. At its core are household-based surveys comprised of child, household and community questionnaires administered every 3 to 4 years with 12,000 children in two age cohorts (born 7 years apart), their households and community members.[12] The pooling of household-based survey data across countries and survey rounds is possible because the rounds occur in the same year and use the same constructs, instruments, questions and variables in each country. The surveys are complemented by longitudinal qualitative research with a sub-sample of the children, as well as school-based surveys with selected children from the Younger Cohort and their peers, while occasional qualitative sub-studies drill down on specific topics, such as young marriage, cohabitation and parenthood.

The research components are intended as far as possible to complement each other so that data generated by each component can be pooled and/or linked with data from the other components. Qualitative data are collected through a variety of semi-structured instruments, with many of the questions mirroring those included in the surveys, and the data are coded using a range of constructs that as far as possible map onto the variables used in the quantitative data. This permits triangulation of evidence, which increases the research's credibility and validity, while also providing greater granularity and completeness of data. For example, the qualitative research offers rich contextual data in narrative and visual formats about children's current circumstances and their perspectives on their lives, permitting insights into the factors underpinning some of the trends observed in the survey data.[13]

In the household-based surveys, variables are given identical designations so that evidence can be compared across age cohorts, survey rounds and countries, permitting analysts to run models and test hypotheses that apply to the full dataset. The large sample size and cross-national analysis of pooled data from a diverse demographic population gives greater statistical power and makes it possible to establish the validity of interpretations derived from country-level evidence. One of the most important data-pooling tasks is the generation of reliable and valid measurements that appropriately depict developmentally salient constructs as they evolve across the life course, as well as participants' changing circumstances and daily activities across the four countries. For example, receptive vocabulary and mathematics have been tested in all countries over time, and there are quite large differences across countries as well as ages, which poses a challenge for comparison. Where there are common items, item response theory (IRT) is used to help strengthen the links across groups and time.[14] To further facilitate harmonisation across countries and survey rounds for the whole dataset, including by external users of the data, Young Lives created a constructed dataset,[15] or panel dataset, which combines selected data from the five survey rounds so that key variables, such as a child's weight, height, and food consumption, can be readily tracked over time.

The Young Lives constructed panel dataset combines sub-sets of selected variables from Rounds 1 to 5 of the household and child surveys. The dataset contains hundreds of variables that are comparable across rounds. The panel format of the dataset classifies data into four pillars: panel information, general characteristics, household characteristics, and child characteristics. The Guide to Young Lives Rounds 1 to 5 Constructed Files is designed to support researchers in their use and analysis of Young Lives data (Briones 2018).

Young Lives' survey instruments are posted on its website to encourage replication of the study's research questions, variables and constructs, with the wider aim of promoting consistency, comparability and use of longitudinal cohort data in LMICs. The instruments have been incorporated into a number of other longitudinal studies. For example, the Gender and Adolescence: Global Evidence (GAGE) study's Ethiopia Baseline Survey includes questions adapted from Young Lives' Round 2 and Round 4 modules on education, paid work, time allocation and social inclusion (Baird et al. 2019). Similarly, Young Lives' labour market and time use questions in Peru inspired a number of questions that were included in the Millennials in Latin America and the Caribbean survey that was administered in Brazil, Chile,

11 The authors aggregated material well-being into a single index using the first component from a Principal Components Analysis, providing proxies for inequalities in living standards. For the second dimension, they used the years of schooling of the most educated parent in the household. The two dimensions were combined into a single measure of social and economic status, using the first Principal Components Analysis component.

12 See https://www.younglives.org.uk/content/household-and-child-survey

13 The qualitative and quantitative data are linked by standardised child, caregiver, household and community identifiers and an organised file-naming system, which enables researchers to easily connect various data types and undertake mixed-methods analysis.

14 See León and Singh (2017) and León (2020) for a discussion of the procedures employed by Young Lives to achieve equivalent cognitive measures across survey rounds and cohorts.

15 See https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=7483

Colombia, El Salvador, Haiti, Mexico, and Paraguay. Data from this survey were also linked with data from the Estudio Longitudinal del Bienestar in Uruguay.[16] The resulting Millennials in Latin America and the Caribbean: To Work or Study? report examines the factors influencing young people's decisions around study and work (Inter-American Development Bank 2018). The analysis, based on both quantitative and qualitative data, found that education access, years of schooling, socio-economic status, young parenthood and family environment were key predictors (Novella et al. 2018). The report is intended to assist evidence-based policy decisions for the benefit of young workers in Latin America and the Caribbean, in the face of significant changes in the labour market.[17]

Most recently, during the current COVID-19 pandemic, Young Lives has harmonised phone survey questions with survey questions developed previously by J-PAL, the Inter-American Development Bank, the World Bank, and the Wellcome Trust.[18] This procedure makes it possible to speed up survey design and administration and to generate more comprehensive, comparable and accessible data, in turn facilitating their wider use. This is essential given the urgency of combatting the multiple effects – both health and economic – of the pandemic.

Data harmonisation challenges

Challenges created by data harmonisation in longitudinal research

- Data harmonisation is complex and time-consuming and requires access to measurement expertise, resources and effective data management.
- The lack of standardisation of instruments, measures and indicators is one of the most significant challenges in cross-study data harmonisation, as is attaining construct equivalence.
- Attempts to harmonise all data toward the lowest common denominator can cause information loss, and disconnection from local meanings and circumstances.

Despite the many advantages, pooling longitudinal LMIC data is a complex and time-consuming process and requires effective data management procedures, close coordination and collaboration between studies and researchers, and access to both appropriate measurement expertise and resources (CLOSER 2020). It is also important to recognise that pooling can be a risk to data quality. For example, arriving at the lowest common denominator to ensure that pooled measures and data are a true representation of common latent constructs can cause the loss of important metadata and disconnection from local meanings and circumstances (CLOSER 2020). Retrospective pooling of longitudinal cohort data involves particular challenges, including the need to mitigate against the potential confounding of cohort, historical, and maturational effects both within and across studies (Curran et al. 2008). Often, data pooling also means adjusting for significant differences in cross-study research objectives.

In order to be amenable to simultaneous analysis, pooled measures of core constructs must hold the same meaning and value across different studies, settings, demographic groups and languages. The more abstract and culturally embedded constructs require considerable work in reconciliation and adaptation to ensure cultural appropriateness, as well as local validation (Dawes 2020). Even with seemingly universalised and relatively concrete constructs the methods and variables used in operationalising them frequently differ. For example, though blood pressure is globally recognised as a key indicator of heart health, the five studies in the COHORTS sample used different instruments and measurements to gather these data, necessitating retrospective harmonisation. Similarly, in high-income contexts, household wealth is frequently measured as the amount of income a household receives. However, in LMICs, where informal employment is more common, households often depend on multiple income sources whose amounts fluctuate considerably over time, so that patterns of consumption or durable assets are commonly regarded as more appropriate indicators of wealth (Howe et al. 2012). Even then, assets can vary widely across contexts in value and type, and their relevance as a measure of wealth may not hold in all settings. For example, many studies focus on reproducible capital such as durable structures or equipment, to the neglect of personal items (such as jewellery or clothing), which can sometimes better capture the intra-household distribution of wealth, or natural resources that may be a more relevant measure of wealth in some rural settings in particular (Emran, Robano and Smith 2014; Rutstein and Johnson 2004).

For Weber et al. (2019), harmonising data retrospectively from 16 longitudinal cohort studies involved several major technical challenges. These included validation results being affected by differences in sampling strategies across studies, the number of items administered for any given child varying considerably across and within cohorts, and items from certain instruments that performed poorly in the model needing to be removed, resulting in the loss of one cohort and of selected data from the remaining cohorts. Researchers in the COHORTS study cite a number of difficulties they encountered, including differences in variable definitions, and in measurement techniques across sites, the different ages of the individuals in the five cohorts and the different ages for which data are available, together with the different time periods the studies covered

16 See http://fcea.edu.uy/estudio-del-bienestar-multidimensional-en-uruguay/108-departamentos/departamento-de-economia/proyectosiecon/estudio-longitudinal-de-bienestar-en-uruguay.html

17 The researchers recommend that public policies focus on interventions related to improving access to relevant and quality skills development, as well as providing accurate information about the job market, in order to strengthen opportunities in education and the workforce for young people (Novella et al. 2018).

18 For more details on the phone survey, see https://www.younglives.org.uk/node/8941

and heterogeneity in the results for some of the analyses, for example, those on body composition (Richter et al. 2012). As a result, they restricted their analyses to those with variables collected consistently across the cohorts, and to ages for which data were available for all (or most) cohorts. Because the various studies collected data from participants at different ages, different outcome variables had to be used for the different cohorts.

Although prospective harmonisation is generally more straightforward than retrospective harmonisation, pooling comparative longitudinal data in a study such as Young Lives, which draws on diverse research methods, data types and units of observation, is both resource and time intensive. As such, despite the good intentions of researchers, not all of the Young Lives data are fully harmonised. For example, due to the challenge of standardising the measurement of household socio-economic status across both survey rounds and contexts, the study employs a multidimensional wealth index,[19] since a multidimensional measure covers more bases, increasing the likelihood that once piloted and validated it will be applicable in all contexts. Nevertheless, cross-country comparisons of household wealth can be misleading. Rather, Young Lives uses this measure to report who is richer and who is poorer relative to national averages within specific countries, with the index providing sufficient coherence to understand the relationship between household wealth and the child outcomes of interest. Similarly, cognitive achievement tests were piloted in Young Lives by diverse panels, to ensure that they were fair to different respondent groups.[20] When an instrument was biased in any way, for example by not cohering effectively with local cultural constructs, it was translated into a national version in a way that retained comparability during the harmonisation and analysis processes (Cueto and León 2012). Nevertheless, full comparability is not always possible. For example, the Peabody Picture Vocabulary Tests (PPVT), a test of children's receptive vocabulary, had to be adapted because it includes images of certain animals that do not exist in all Young Lives sites.

Language differences can be an obstacle to pooling comparative data. For example, in the PPVT the words increase in difficulty towards the top end of the test. While this may be an effective way of distinguishing children's command of vocabulary in many Indo-European languages, such as Spanish, it does not work so well in Vietnamese, where complex terms are broken down into different words that can be easily understood by everyone. Young Lives finds that adapting instruments appropriately to minority local languages can be especially time consuming and requires careful training as well as use of interpreters (Cueto and León 2012). If translation obstacles are not properly addressed, the risk is that participants who speak minority languages may be left out of or misrepresented in data analysis. Further, not considering the cultural contexts of questions may mean that data on certain subjects are of higher quality in one country than they are in other contexts, leading to difficulties when comparing across country contexts.[21]

Data linkage opportunities to expand and strengthen research

Opportunities created by data linkage in longitudinal research

- Data linkage can enhance longitudinal datasets, and create interoperability between them, which allows researchers to access expanded sets of variables, thus increasing the completeness and usefulness of data.
- Linkage increases opportunities for innovation and may allow more uptake and policy influencing opportunities with national authorities, as researchers can ask new questions of the linked dataset without disrupting the original longitudinal survey design.
- Linking longitudinal survey data with administrative data can serve as an alternative to impact evaluations in contexts where randomised controlled trials are too challenging or expensive to implement.

Due to cost and the need to keep attrition to a minimum, there will always be limitations in longitudinal research when it comes to the number of questions that can be asked in a questionnaire. Also, the requirement that instruments and measures be repeated across data rounds in order to build an effective panel dataset means that longitudinal studies cannot easily respond to changing contexts or new information needs. However comprehensive the initial design, it is impossible to anticipate at the outset all of the lines of enquiry that will subsequently emerge as important (Boyden and Walnicki 2020). Data linkage has the huge advantage of making it possible to expand the scope of research by supplementing survey data at limited additional cost and without increasing the burden on survey participants. It can also facilitate validation of survey data and may provide information that might be difficult to obtain through direct interviews. It may also reduce recall bias for complex questions (e.g. recalling weekly benefits received from a support programme over the course of more than a year). Data linkage may be conducted between two distinct data sources or within a single dataset to identify multiple entries for an individual or record unit.

Given the frequent attention of cohort studies to the development and well-being outcomes of exposure to varying environmental factors, community-level data are clearly enormously important to this kind of research. With

19 Developed initially by the World Bank and Macro International (Briones 2017).

20 For more information about fairness in the Young Lives cognitive achievement tests, see Cueto et al. (2009).

21 Research instruments are created in English, and then translated into each study country's main national language using a process of double-translation (Dawes 2020).

the objective of furthering the development of data linkage methodologies and focusing specifically on how these can be used to benefit cohort studies, the UK-based Avon Longitudinal Study of Parents and Children (ALSPAC) team is engaged in scoping data science and governance issues relating to linking geospatial data.[22] The team deploys an innovative technique in which they link study data to residential and neighbourhood data, using postcodes. Although their model is for application to a specific region of the UK, the intention is to develop a methodological template for use globally. Their Enhancing Environmental Data Resources in Cohort Studies: ALSPAC exemplar (ERICA) programme in particular has considerable potential for linking cohort and environmental data. Through ERICA the team developed a geocoded database for the ALSPAC cohort, including participant residential addresses across three generations. Use of the Utilisation of the Algorithm for Generating Address Exposures (ALGAE) software has made it possible for the study to link location-based data to participants to calculate individual-level exposures at key life stages to stressors in the built and natural environments, such as air and noise pollution and extreme temperatures.

Recognising the risks to child health, growth and development in areas with poor public health infrastructure and low levels of hygiene that are also exposed to extreme weather events, investigating the community-level determinants of child nutrition and development has been a major priority for Young Lives. The survey gathers basic geospatial data on services, infrastructure and income sources for the sites and communities in which the young people live. However, these data are far less detailed and comprehensive than information gathered at the individual, household and school levels. The team explored the feasibility of gathering local meteorological data either continuously or periodically during survey rounds but concluded that this would not be possible, mostly because of the technical complexity and cost. Linking Young Lives household and community data with meteorological data proved very effective in overcoming this limitation. In a study of the association between nutrition at different periods from conception to middle childhood and cognitive achievement in early adolescence, precipitation and temperature data from the Global Climate Database of the University of Delaware (UDEL) were linked to the communities where the children were residing in each round using the communities' geographical coordinates (Georgiadis 2017). A conceptual framework was developed that delineates the channels through which child health impacts cognitive development at different ages from conception to middle childhood and uses exogenous variation in nutritional status arising from weather shocks. It was found that even though there was an association between extreme weather events and the prevalence of infectious diseases and undernutrition among infants and children has negative implications for child growth and cognitive development, there is important scope for remediation in later stages of childhood (Georgiadis 2017).

Fan and Porter (2020) used the same linked data, and information on siblings of the Young Lives Younger Cohort children to study how variations in early life conditions (proxied by weather) caused sibling differences in health and cognitive achievement. They found that parents then attempt to compensate by investing more in the disadvantaged sibling, to the extent of their limited resources. Another Young Lives study, in Peru, tested the hypothesis that particularly low temperatures (below long-term averages) experienced in utero and during the first 36 months of life adversely affect nutritional status (proxied by height-for-age), this being a major channel influencing vocabulary achievement and self-esteem, outcomes that have been shown to help explain later differences in labour market results (Sánchez 2018). The empirical strategy used differences in exposure to temperature levels across children within clusters, generated by differences in date of birth, at the month precision, in areas where frosts are widespread. Responses to the community questionnaires revealed that around half of the Young Lives households in Peru situated at high altitudes had been affected by frosts. Study data were linked with temperature data provided by Peruvian Institute of Meteorology weather stations located either in the selected clusters or in nearby clusters (Sánchez 2018). It was found that while exposure to unusual weather variations can affect child development, recovery is possible in some dimensions and the impact can vary by gender.

Services and programmes are also a key contributor to young people's development and well-being. Although longitudinal observational studies are not generally designed to examine the effects of specific interventions, data linking has been used to trace which Young Lives children and households are accessing which services and programmes and with what implications for their development. In this sense, linking longitudinal survey data with administrative data can serve as an alternative to impact evaluations in contexts where randomised-controlled trials are too challenging or expensive to implement. As such, linking can be an extremely important tool in policy and programme design. The Young Lives pro-poor sampling design is a valuable attribute when linking study data to administrative data derived from social protection programmes.[23] In one case, rainfall and weather data were linked with nutritional shock data and data from the National Rural Employment Guarantee Scheme, a public works programme in India, to examine how social protection policies impact people who have experienced drought shocks (Dasgupta 2017). By combining administrative data about drought shocks and geospatial data with data on the health outcomes of Young Lives respondents, the study found that the social protection programme can buffer against the negative health impacts of drought shocks. However, the effect of the buffer was found to vary by population subgroup and did not make up for long-term health deficiencies.

22 See http://www.bristol.ac.uk/alspac

23 The design means that households in poorer regions of the four study countries are deliberately over-sampled.

In two recent studies, researchers combined Young Lives data with other sources in order to understand the impact of the extended school-day reform Jornada Escolar Completa (JEC) in Peru. Sánchez and Favara (2019) used DHS and Young Lives data to assess the impact that the JEC had on factors that predict teenage pregnancy. Agüero et al. (2021) investigated the programme's impact on learning outcomes as well as the mechanisms for impact. Data from the Younger Cohort in Peru were linked to an administrative dataset at school level in order to identify which children were attending JEC schools. The study also used the 2015 Evaluación Censal de Estudiantes de Secundaria (ECE), the School Census, Semáforo Escuela (a data system used by the Ministry of Education to monitor the delivery of educational services) and the National Survey of Teachers (Encuesta Nacional a Docentes, ENDO). The linkage process was lengthy but straightforward and involved matching the administrative dataset with Young Lives Round 5 survey data using the names of the educational institutions. Linking the datasets enabled researchers to determine that young people who participate in the programme experience, on average, improved school performance, and increased aspirations and socio-emotional competencies.

The benefit of employing both data linking and data pooling across different research components within a study is seen in the Young Lives school-based research, which was introduced in order to help explain how different school factors shape children's outcomes. The intention was to examine the schools attended by the Younger Cohort, while also assessing the children's attainment in class, a design that was applied in Ethiopia (2010) and India (2010–11) (Boyden and James 2014). Sampling procedures for the school surveys vary by country, but each survey visits either the schools attended by a sub-sample of children from the household sample, or a sample of the schools located within the geographical boundaries of the Young Lives site. Data are collected at the school, principal, class, teacher and pupil levels and include a number of innovative features, such as teacher professional knowledge (Moore and Rossiter 2018). Whenever possible using instruments also administered in the household surveys, child outcomes focus on cognitive development (language and mathematical skills) and children's transferable skills, for example problem solving and critical thinking (Iyer and Azubuike 2017). Children's attainment in these domains is measured in relation to a specified grade – usually that in which the majority of the Younger Cohort are enrolled.

The school surveys make it possible in some contexts and survey rounds to link child and household data from the regular survey rounds with school data and, for children sampled in both surveys, to map their learning against their background characteristics and the features of the schools they attend. In this way, researchers can contextualise and explain trends observed in the household data, such as how in India children from poorer households and children with less-educated mothers are more likely to attend ineffective schools (Rolleston and Moore 2018). This research contributes to policies to improve school effectiveness, combat inequality in education and break the intergenerational transmission of poverty (Morrow 2017).

Data linkage challenges

Challenges created by data linkage in longitudinal research

- Data from external datasets may be inaccessible, incomplete, irrelevant or poor quality.
- There are few templates for linking data in LMICs, with most methodologies centring on high-income contexts where the specificity of research instruments, number of datasets and opportunities for linking are far greater.
- Data linking can involve a number of ethical challenges that need to be addressed.

Data linkage in LMICs shares a number of the same methodological constraints and limitations as those affecting data pooling and can entail additional bureaucratic, technical and ethical challenges. There are few templates for linking data in LMICs, with most existing methodologies centred on high-income contexts where the specificity of research instruments, availability of metadata, number of datasets and opportunities for linking are far greater. For example, efforts to administer the ALSPAC model linking geospatial and meteorological data with longitudinal survey data in LMICs would likely confront major constraints, due to data shortfalls, the lack of comprehensive postcodes, especially in rural areas, and other challenges. When attempting to link at the individual level, things become even more complex, when individuals do not have any identification number, birth dates are often misremembered, and the same name can be both common, and spelled in several different ways. In the Young Lives Peru dataset, the team collected a national identifying number (DNI) which is commonly used in administrative datasets, but no such identifier exists in Ethiopia, for example.

The significant limitations in the quality, availability and utility of administrative data in many LMICs is one of the most significant constraints. For instance, the Center for Global Development review of administrative data in the education sector found that out of 133 LMICs, 61 had no available data and 43 had data only at the national level (Rossiter 2020). Of the 29 countries that did have disaggregated data, the majority were in PDF or non-downloadable format and only 16 countries provide information from student assessments. When administrative records do not exist for all survey respondents, this may bias analyses using the administrative data if survey participants for whom the data are not available differ from those for whom they do exist, or who simply cannot be matched due to uncertainty about their identity in the dataset.

Although opening up exciting opportunities for enquiring into the contribution of different school characteristics to children's learning outcomes, the introduction of school surveys into Young Lives posed a number of specific data pooling and linking challenges. Options for aligning the datasets are in practice limited, a problem that confronts many data linking efforts. While much of the content of Young Lives school surveys necessarily varies across

countries, in keeping with national priorities in education policy, consistency of design and instruments across research rounds and countries is crucial for the household-based surveys. More challenging still is the significant divergence in children's levels of attainment at school, both within and across country contexts. When establishing the school surveys, the Younger Cohort was spread widely across schools and classes, and many children were in grades that were lower than the one expected for their age, or had dropped out. Thus, their schooling was not in effect comparative, so that switching an age cohort into an effective school cohort resulted in inconsistencies in the numbers of children per class and school. To ensure continuity in the child-level panel and at the same time create school-based datasets, the design was adjusted to include children at both class and school levels who were not in the household sample. As a consequence, in some sites and research waves there are insufficient children from the household sample included in the school surveys to permit pooling or linking of data across the full Young Lives datasets.

Data linking may be less complex conceptually and have different technical challenges to data pooling but it presents major ethical concerns. First, marginalised populations such as workers in the informal sector and those who are internally displaced by conflict are often excluded from government datasets. Data linking may therefore perpetuate biases in administrative data and further exclude the already marginalised. Second, data linking is generally undertaken retrospectively, and may not have been part of the original study objective or design. This raises questions about the need for and feasibility of renewing participants' informed consent given that the linked data are often intended for use in novel ways and for different ends (Morrow 2013). Third, data linking necessitates the use of personal data – names, addresses, GPS coordinates – so that a study sample can be connected with records on that same sample in an external dataset. The risk here is that data linking may breach the confidentiality and anonymity of study participants. For example, geospatial data are often open access and necessarily entail location identifiers, with GPS coordinates providing a high level of specificity. Administrative datasets do not always comply with high standards in relation to the protection of personal data. Thus, in some datasets individual and household identifiers and personal details, such as sources and levels of income, are readily accessible on the internet.

The ALSPAC team is engaged in scoping data science and governance issues relating to linking spatial data with a view to addressing some of these concerns. Their projects include the development and assessment of statistical approaches to de-identify and anonymise data and designing anonymised protocols for data linkage. One of the most common approaches to this problem is to provide for a degree of separation between linkage and analysis processes so that identifiable data are accessed only by those who conduct the linkage, while those involved in the analysis only access fully anonymised data. For example, Young Lives has put in place a protocol that strictly limits access to participants' personal data to current staff, with assistance given to selected external researchers to enable them to link anonymised Young Lives data with other datasets. While this protocol protects the identities and locations of study participants, it does impede assessment data linkage quality.

Summary

Longitudinal data are a vital asset in the advancement of science and policy and such data have the potential to make an especially valuable contribution in LMICs. However, longitudinal research encounters major limitations in many LMICs due to small sample sizes, constraints to the generalisability of findings, the limited number of instruments that have been fully validated in these contexts and shortfalls in high-quality, accessible data. Moreover, the reliance of longitudinal research on repetition of instruments and measures across data rounds means that it cannot readily respond to changing contexts or new information needs. This paper has argued that data linking and harmonisation increase the usability of existing longitudinal data in LMICs, extend their scientific reach and policy and programme impact at far lower cost than is possible by gathering new data. Data harmonisation expands sample size and heterogeneity, thereby heightening statistical power and the generalisability of findings across settings. Data linking enlarges the number of variables available on specific samples, increasing insights and allowing new lines of enquiry into participants' circumstances and outcomes.

Beyond this, harmonisation involves the reconciliation of constructs, variables, indicators and measures across longitudinal studies, and helps to improve the consistency, comparability and integration of such data in LMICs, in turn providing an opportunity for future harmonisation to be undertaken prospectively. At the same time, recognising the current constraints of quality, access and relevance, as the demand for evidence-based policy and programmatic decisions grows, governments will likely gather more data, further increasing the opportunities for linking longitudinal data with administrative data.

That said, even though potentially powerful tools, data harmonisation and linkage have their limitations and challenges. Both can be time-consuming, resource intensive, and technically and logistically complex, and harmonised data can become so generic as to have limited applicability to specific groups and contexts. By creating networks of researchers who are engaged in longitudinal research and ensuring greater involvement and ownership of such research by local scholars, collaborative data pooling initiatives can help overcome some of these difficulties. There can be significant ethical concerns with data linking in particular. Special attention must be paid to ensuring linked data do not breach participants' anonymity and confidentiality or exclude the already marginalised. Researchers must ensure that such challenges as discussed in this paper can be mitigated, and participants' data are both protected and utilised in the most expansive, impactful way possible.

References

Adair, L., C. Fall, C. Osmond, A. Stein, R. Martorell, M. Ramirez-Zea, H.M. Singh Sachdev, D. Dahly, I. Bas, S. Norris, L. Micklesfield, P. Hallal, and C. Victora (2013) 'Associations of Linear Growth and Relative Weight Gain During Early Life with Adult Health and Human Capital in Countries of Low and Middle Income: Findings From Five Birth Cohort Studies', *The Lancet* 382.9891: 525–34.

Agüero, J.M., M. Favara, C. Porter, and A. Sánchez (2021) *Do More School Resources Increase Learning Outcomes? Evidence from an Extended School-Day Reform*, IZA Institute of Labor Economics Discussion Paper No. 14240, Bonn: IZA Institute of Labor Economics.

Alkire, S., U. Kanagaratnam, and N. Suppa (2020) *The Global Multidimensional Poverty Index (MPI) 2020*, OPHI MPI Methodological Note 49, Oxford: OPHI.

Alkire, S., S. Kovesdi, C. Mitchell, M. Pinilla-Roncancio, and S. Scharlin-Pettee (2020) *Changes over Time in the Global Multidimensional Poverty Index*, OPHI MPI Methodological Note 50, Oxford: OPHI.

Apablaza, M., and G. Yalonetzky (2013) *Decomposing Multidimensional Poverty Dynamics*, Young Lives Working Paper 101, Oxford: Young Lives.

Baird, S., J. Hicks, N. Jones, J. Muz, and the GAGE consortium (2019) 'Ethiopia Baseline Survey 2017/2018', http://doc.ukdataservice.ac.uk/doc/8597/mrdoc/pdf/8597_gage_ethiopia_baseline_survey_cr.pdf (accessed 26 April 2021).

Boyden, J., and D. Walnicki (2020) 'Opportunities, Challenges And Strategies In Generating And Governing Longitudinal Data – Learning From Two Decades At Young Lives', Oxford: Young Lives.

Boyden, J., and Z. James (2014) 'Schooling, Childhood Poverty and International Development: Choices and Challenges in a Longitudinal Study', *Oxford Review of Education* 10.1: 10–29.

Briones, K. (2018) *A Guide to Young Lives Rounds 1 to 5 Constructed Files*, Young Lives Technical Note 48, Oxford: Young Lives.

Briones, K. (2017) *'How Many Rooms Are There in Your House?' Constructing the Young Lives Wealth Index*, Young Lives Technical Note 43, Oxford: Young Lives.

CLOSER (Cohort and Longitudinal Studies Enhancement Resources) (2020) 'Preparing for the Future II: International Approaches to Challenges Facing the Longitudinal Population Studies', London: Economic and Social Research Council.

CLOSER (2018) 'CLOSER Takes Data Discoverability Worldwide in New International Project', https://www.closer.ac.uk/news-opinion/news/closer-international-project-launch (accessed 26 April 2021).

Cueto, S., and J. León (2012) *Psychometric Characteristics of Cognitive Development and Achievement Instruments in Round 3 of Young Lives*, Young Lives Technical Note 25, Oxford: Young Lives.

Cueto, S., J. Leon, G. Guerrero and I. Munoz (2009) *Psychometric Characteristics of Cognitive Development and Achievement Instruments in Round 2 of Young Lives*, Young Lives Technical Note 15, Oxford: Young Lives.

Curran, P., A. Hussong, L. Cai, W. Huang, L. Chassin, K. Sher, and R. Zucker (2008) 'Pooling Data from Multiple Longitudinal Studies: The Role of Item Response Theory in Integrative Data Analysis', *Development Psychology* 44.2: 365–80.

Das, J., A. Singh, and A. Yi Chang (2020) *Test Scores and Educational Opportunities: Panel Evidence from Five Developing Countries*, RISE Working Paper Series 20/040, Oxford: RISE.

Dasgupta, A. (2017) Can the Major Public Works Policy Buffer Negative Shocks in Early Childhood? Evidence from Andhra Pradesh, India', *Economic Development and Cultural Change* 65.4: 767-804.

Dawes, A. (2020) 'Measuring the Development of Cognitive Skills Across Time and Context: Reflections from Young Lives', Insights Report, Oxford: Young Lives.

Emran, M.S., V. Robano, and S.C. Smith (2014) 'Assessing the Frontiers of Ultrapoverty Reduction: Evidence from Challenging the Frontiers of Poverty Reduction/Targeting the Ultra-poor, an Innovative Program in Bangladesh', *Economic Development and Cultural Change* 62.2: 339–80.

Fall, C.H., C. Osmond, D.S. Haazen, H.S. Sachdev, C. Victora, R. Martorell, A.D. Stein, L. Adair, S. Norris, L. Richter, and COHORTS investigators (2016) 'Disadvantages of Having an Adolescent Mother', *Lancet Global Health* 4: 787–88.

Fall, C.H., H.P.S. Sachdev, C. Osmond, M.-C. Restrepo-Mendez, C. Victora, R. Martorell, A.D. Stein, S. Sinha, N. Tandon, L. Adair, Basl, S. Norris, L. Richter, and the COHORTS Group (2015) 'Associations Between Maternal Age at Childbirth and Child and Adult Outcomes in the Offspring: A Prospective Study in Five Low-income and Middle-income Countries', *Lancet Global Health* 3: 366–77.

Fan, W., and C. Porter (2020) '*Reinforcement or Compensation? Parental Responses to Children's Revealed Human Capital Levels in Ethiopia', Journal of Population Economics* 33.1: 233–70.

Georgiadis, A. (2017) *The Sooner The Better But It's Never Too Late: The Impact of Nutrition at*

Different Periods of Childhood on Cognitive Development, Young Lives Working Paper 159, Oxford: Young Lives.

Howe, L.D., B. Galobardes, A. Matijasevich, D. Gordon, D. Johnston, O. Onwujekwe, R. Patel, E.A. Webb, D.A. Lawlor, and J.R. Hargreaves (2012) 'Measuring Socio-economic Position for Epidemiological Studies in Low- and Middle-income Countries: A Methods of Measurement in Epidemiology Paper', *International Journal of Epidemiology* 41.3: 871–86.

Inter-American Development Bank (2018) 'Millennials in Latin America and the Caribbean: To Work or Study?', https://publications.iadb.org/publications/english/document/resumen-ejecutivo-millenials-ing-web.pdf (accessed 26 April 2021).

Iyer, P., and O.B. Azubuike (2017) *Young Lives School Surveys 2016–17: The Design and Development of Transferable Skills Tests in India and Vietnam*, Young Lives Technical Note 42, Oxford: Young Lives.

León, J. (2020) *Equating Cognitive Scores Across Rounds for Young Lives in Ethiopia, India, Peru and Vietnam*, Young Lives Technical Note 51, Oxford: Young Lives.

León, J., and A. Singh (2017) *Equating Test Scores for Receptive Vocabulary Across Rounds and Cohorts in Ethiopia, India and Vietnam*, Young Lives Technical Note 40, Oxford: Young Lives.

Moore, R., and J. Rossiter (2018) *Young Lives School Surveys, 2016-17: The Design and Development of Teacher Measures for Use in Ethiopia, India and Vietnam*, Young Lives Technical Note 44, Oxford: Young Lives.

Morrow, V. (2017) 'A Guide to Young Lives Research', Oxford: Young Lives.

Morrow, V. (2013) 'Practical Ethics in Social Research with Children and Families in Young Lives: A Longitudinal Study of Childhood Poverty in Ethiopia, Andhra Pradesh (India), Peru and Vietnam', *Methodological Innovations Online* 8.2: 21–35

Novella, R., A. Repetto, C. Robino, and G. Rucci (2018) 'Millennials in Latin America and the Caribbean: To Work or Study?' Washington, DC: Inter-American Development Bank.

OPHI (2020) 'Global Multidimensional Poverty Index 2020 - Charting Pathways out of Multidimensional Poverty: Achieving the SDGs', https://ophi.org.uk/wp-content/uploads/G-MPI_Report_2020_Charting_Pathways.pdf (accessed 26 April 2021).

Richter, L., C. Victora, P. Hallal, L. Adair, S. Bhargava, C. Fall, N. Lee, R. Martorell, S. Norris, H. Sachdev, A. Stein, and the COHORTS Group (2012) 'Cohort Profile: The Consortium of Health-Orientated Research in Transitioning Societies', *International Journal of Epidemiology* 41: 621–26.

Rolleston, C., and R. Moore (2018) 'Young Lives School Survey, 2016-17: Value-added Analysis in India', Oxford: Young Lives.

Rossiter, J. (2020) 'Link It, Open It, Use It: Changing How Education Data Are Used to Generate Ideas', Washington, DC: Center for Global Development.

Rutstein, S.O., and K. Johnson (2004) 'DHS Comparative Reports 6: The DHS Wealth Index', Calverton, MD: ORC Macro.

Sánchez, A. (2018) *Early-life Exposure to Weather Shocks and Human Capital Accumulation: Evidence from the Peruvian Highlands,* Young Lives Working Paper 178, Oxford: Young Lives.

Sánchez, A., and M. Favara (2019) *Consequences of Teenage Childbearing in Peru: Is the Extended School-day Reform an Effective Policy Instrument to Prevent Teenage Pregnancy?* Young Lives Working Paper 185, Oxford: Young Lives.

UNICEF (n.d.) 'Global Longitudinal Research Initiative', https://www.unicef-irc.org/files/upload/documents/About%20GLORI.pdf (accessed 26 April 2021).

Victora, C.G., L. Adair, C. Fall, P.C. Hallal, R. Martorell, L. Richter, H.S. Sachdev, and the Maternal and Child Undernutrition Study Group (2008) 'Maternal and Child Undernutrition: Consequences for Adult Health and Human Capital, *The Lancet* 371.9609: 340–57.

Weber A.M., M. Rubio-Codina, S.P. Walker, et al. (2019) 'The D-score: A Metric for Interpreting the Early Development of Infants and Toddlers Across Global Settings', *BMJ Global Health* 4.6: e001724

The project

The Methodological Lessons and Learning in Longitudinal Research project, funded by the ESRC, aims to strengthen capacity and effectiveness in the conduct of longitudinal research in low- and middle-income countries, while also contributing to a growing community of practice. Through this project, Young Lives is reflecting on its experience and past practices over 20 years of research and working with experts in longitudinal research to share learning and support researchers running large-scale longitudinal studies in international development and related fields.

The authors

Professor Emeritus Jo Boyden was based at Oxford University's Department of International Development and was Director of Young Lives from 2005 to 2019. Her research has mainly focused on child labour, children and political violence, and childhood poverty – particularly in bringing together academics, practitioners and policymakers to develop effective models and methods for supporting children, their families and their communities in situations of adversity.

Deborah Walnicki joined the Young Lives team in 2018 as a Qualitative Research Assistant, and worked as Research Consultant until April 2021. Currently, she is the Mapping Coordinator for the Evidence for Gender and Education Resource based at the GIRL Center at the Population Council, where her work focuses on girls' education in low- and middle-income countries. Deborah holds an MPhil in Development Studies from the University of Oxford.

Acknowledgments

This paper was developed in consultation with Alan Sánchez, Caine Rolleston, Fanni Kovesdi, Jack Rossiter, and Prerna Banati, experts on the processes of data harmonisation and data linkage, particularly in longitudinal studies in low- and middle-income countries.

We would like to thank Gina Crivello, Georgina Fensom, and Catherine Porter, who kindly reviewed this report and offered valuable inputs. In addition, we are grateful to Adam Houlbrook for copyediting and Garth Stewart for the design of this report.

This Insights Report was funded by a grant from the Global Challenges Research Fund (GCRF), as part of the two-year project – Methodological Lessons and Learning in Young Lives, funded by the Economic and Social Research Council (ESRC). The support of the ESRC is gratefully acknowledged.

GCRF
Global Challenges Research Fund

UKRI Economic and Social Research Council

Young Lives, Oxford Department of International Development (ODID)
3 Mansfield Road, Oxford OX1 3TB, UK

www.younglives.org.uk
Tel: +44 (0)1865 281751 • Email: younglives@qeh.ox.ac.uk • Twitter: @yloxford

Young Lives