Young Lives

Opportunities, Challenges and Strategies in Generating and Governing Data – Learning from Two Decades of Research at Young Lives

Introduction

Since 2001, Young Lives has been studying the development and well-being of 12,000 children growing up in Ethiopia, India, Peru and Vietnam through mixed-methods, comparative, longitudinal research.[1] The research aims to identify the determinants and outcomes of child poverty and explain how policies and programmes can best break inter-generational poverty cycles and reduce inequalities. Data governance is a central element of the study, and strong coordination and collaboration between the study partners ensure that data collected from the four countries are managed effectively. Young Lives' data governance policies focus on two main priorities: maintaining data quality and research rigor, and maximising data use by researchers and policy stakeholders globally. The first priority centres on accountability to research participants and maintaining effective standards and procedures in research design, data collection and data management; the second on data democracy, data discoverability and development of capacity in data analysis.

1 The Young Lives sample was recruited in 2001 and the first survey was conducted in 2002. In India, the Young Lives study is conducted in the states of Andhra Pradesh and Telangana.

This summary of the 'Opportunities, Challenges and Strategies in Generating and Governing Data – Learning from Two Decades of Research at Young Lives' report briefly outlines the Young Lives research design and its implications for the choice of instruments, data gathering and data management. It also reviews the technological developments and operational considerations in survey administration and data management, as well as data democratisation and discoverability. The summary reflects on the challenges Young Lives has faced and aims to offer insights for researchers, programme managers and data managers involved in large-scale longitudinal cohort studies in low- and middle-income countries (LMICs). This summary was written by Deborah Walnicki. The full report by Deborah Walnicki and Jo Boyden is available here, detailing acknowledgements, photo credits and references.

Study design

Young Lives employs a prospective, multidisciplinary, mixed-methods design featuring comparative longitudinal cohort research across four countries, pro-poor sampling, and a multidimensional view of human development. The hybrid model incorporates longitudinal and cross-sectional research involving four key inter-linked components, each one involving a distinct design, with distinct research methods, units of observation and analysis at individual, household, community and school levels. These components work together iteratively to enable triangulation of findings and allow topics arising in one component to be further probed in the others.

The research components comprise household-based surveys, longitudinal qualitative research, school-based surveys and qualitative sub-studies on specific topics, such as paid work, and young marriage, cohabitation and parenthood (Figure 1).

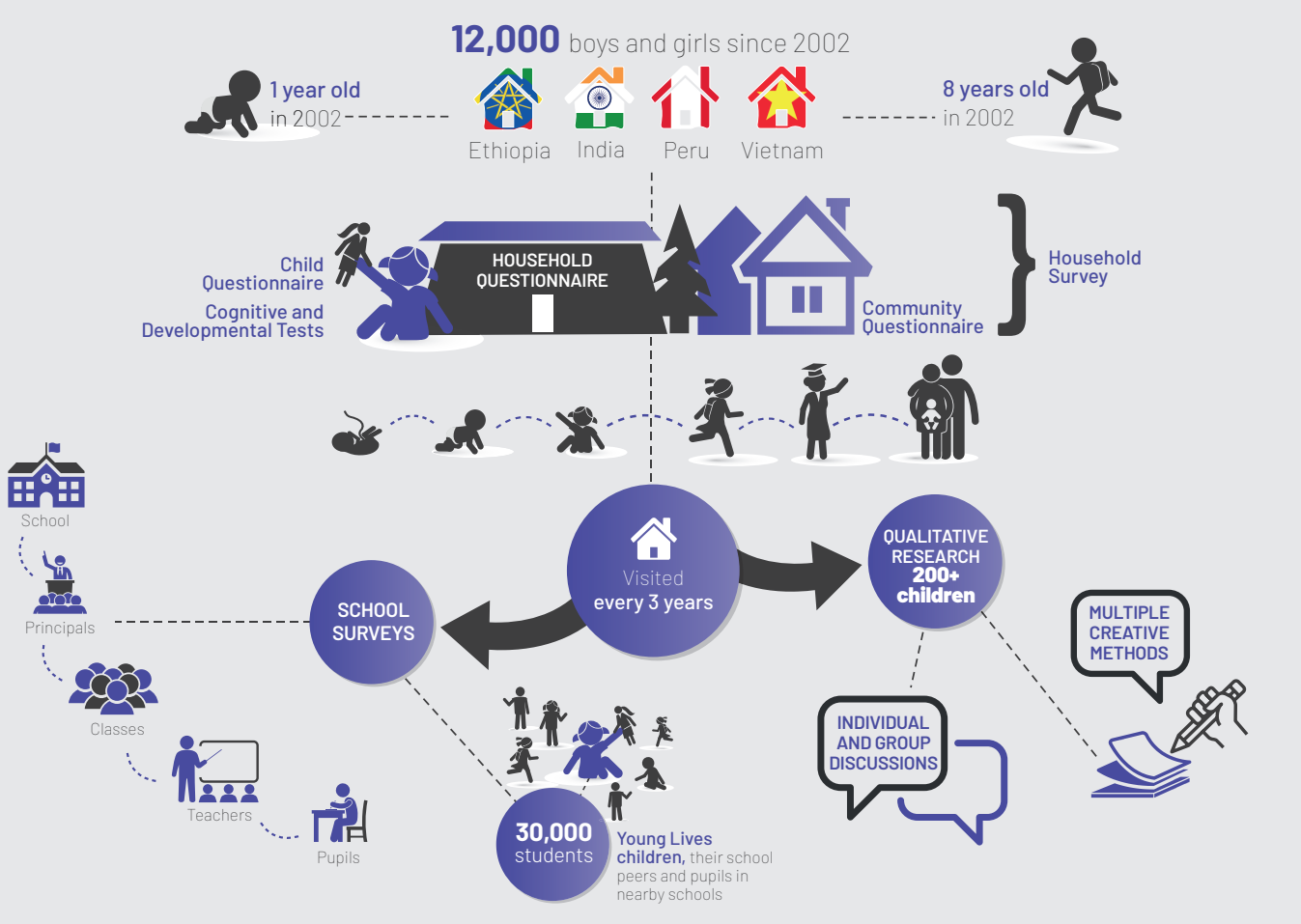
The research design strives to be consistent over time, yet flexible enough to respond to changing environments, opportunities for innovation and aging respondents. It is based on a mix of instruments created specifically for the study and measures originally developed for administration in high-income settings that have been adapted to remain relevant to children growing up in LMICs, and applied across different languages, socio-cultural and economic groups, and in diverse contexts (Dawes 2020).

Implications of research design for instrument development

Longitudinal design

Retention of the same core content and constructs across data rounds is an important feature of the longitudinal qualitative and survey research as this permits direct comparison between data rounds and country samples, as well as allowing the linking of data generated through the different research components. Nevertheless, while striving to maintain continuity of constructs and tools over time, Young Lives has had to adjust instruments and measures to ensure their relevance for respondents' ages, developmental phase, and circumstances. Whenever necessary, instruments are also adapted to improve their performance. As a multipurpose study, Young Lives data allow for multiple lines of enquiry. Even so, a key limitation of longitudinal research is that however comprehensive the initial design, it is impossible to anticipate at the outset all the lines of enquiry that will subsequently emerge as important.

Figure 1: *Young Lives research components*



Hybrid, multi-component research design

Despite the constraints associated with maintaining conceptual and instrument continuity in longitudinal research, it is sometimes possible to draw on the emergent evidence to develop new lines of enquiry and introduce new research modules and elements. In Young Lives, the qualitative research and school components were later additions to the household panel. Although introducing new research modules and linking data to household survey data was complicated, these additions offered new insights, for example on the link between diverse school characteristics and children's learning outcomes, and informed subsequent data collection. Young Lives works continuously to integrate the mixed-methods data at a conceptual level, including by using qualitative data to refine questionnaires. For example, knowledge gained in the qualitative research about adolescents' migration and mobility for school and work was used in the development of a new migration module in the fourth survey round.

Comparative research

Young Lives' comparative design, together with the sociocultural diversity of country samples and their dispersion across rural and urban locations, means undertaking research in multiple languages and with varied population groups who do not necessarily share equivalent concepts and whose experience, worldview and levels of education are disparate. Ensuring that all household- and school-based questionnaires and tests are reliable, valid and an equivalent measure of the skills, attitudes and traits in all population groups is challenging (Attawell 2004; Dawes 2020).

Working across multiple languages and contexts: a view from secondary school surveys

> "Most people say the main issue with surveys in schools is getting official permissions – Young Lives is lucky in that the country teams have excellent relationships with officials and it is already well known as a study collecting different kinds of data, so this was something we didn't have to contend with too much. In terms of practicalities, the translation, piloting of different language versions, and further checking of translation of the survey tools ended up being one of the biggest things to consider, particularly in Ethiopia where we had to have these in many languages. But it wasn't a challenge as such, just something which took a long time!"

(Researcher, school survey team, Oxford)

Key lessons from Young Lives about research design and instrument development

- Though complex to administer, a multidisciplinary, mixed-methods study design that combines longitudinal with cross-sectional research allows effective triangulation of findings and provides greater granularity of data, while also facilitating new lines of enquiry without overburdening respondents.
- Longitudinal research with children needs to be designed flexibly in order to ensure as much consistency as possible in constructs and measures administered across rounds, while at the same time, adapting instruments to respond to children's developmental changes, age-related responsibilities and any shifts in their external circumstances.
- Most instruments and measures administered with children are designed for application in high-income countries or with adults. Consequently, research with children in LMICs needs to take account of the many aspects of their lives and development that are not typically provided for in these tools, particularly their social and economic roles and responsibilities, their transitions through childhood and beyond, and their pro-social and life skills.
- In research administered across different languages and with different sociocultural groups, it is essential to ensure that instruments are reliable, valid and show measurement equivalence in constructs between groups. This calls for considerable attention to pre-piloting and piloting of tools at each research wave, as well as the deployment of translators familiar with the local vernacular and idioms, and the possible need for double translation. Piloting and pre-piloting is most effective when undertaken by mixed teams comprised of researchers, data managers and experienced field supervisors.
- The frequency and timing of data rounds should, as far as possible, reflect the study's goal and its theory concerning the factors likely to affect change and stability in children's development over time. Schedules must also take account of funding and logistical constraints to implementing research in LMICs, as well as adjusting for children's institutional and social transitions and aiming to minimise attrition.

Cohort maintenance and attrition

Sample attrition is one of the most significant threats to longitudinal studies as it can seriously undermine the feasibility of statistical analysis and the authority of the research and introduce bias into findings. Challenges with cohort maintenance in Young Lives included keeping in contact with the dispersed sample over time and respondent migration. Cohort maintenance has been a major success in Young Lives due to efforts to limit the frequency of data rounds and careful tracking of respondents, including phone-based tracking. Country teams are key to this success. Their focus on relationship building and research reciprocity, in addition to the technical aspects of fieldwork, from training, tracking, piloting and data gathering, have all contributed to effective cohort maintenance.

What has helped reduce attrition? A view from India

> "First, continuation of the same field supervisors across survey rounds has had huge advantages. Second, gender matching in who administers the survey; that is, women speaking to women and girls, and men speaking to men and boys. Third, contact with the supervisor between survey rounds; sometimes families contact the supervisors with questions so it is important that the supervisors are available to respond. Fourth, maintaining up-to-date addresses and keeping contact information for someone else in their family in case [participants] migrate ... Finally, research reciprocity, in kind or cash … had it not been there, we can definitely say there would have been greater attrition. It's this combination of strategies that's helped."
>
> (Field coordinator, India)

Cohort maintenance is the responsibility of study-country principal investigators, who work closely with survey coordinators and supervisors in planning and supervising tracking, data collection and data entry, and recruitment and training of enumerators. Training ensured that enumerators understood the research objectives, the content, logic, structure of the research instruments and the strategies for reducing attrition. The main strategies for maintaining the sample and reducing attrition were: careful recruitment and extensive training of field teams; whenever possible, employing the same field supervisors and enumerators across rounds so that they are familiar with the research and the families, communities and authorities; tracking; keeping instrument length and frequency of data rounds to a minimum; and research reciprocity.

Research reciprocity, aimed at giving something back to the children, families and communities for their participation in the research, takes varying forms. In Ethiopia it has mainly involved sharing findings and giving advice to local officials at a regional level, whereas in Peru, Young Lives has shared accessible summaries of research findings with participants in specially tailored information booklets, so that children and their families can understand the study's purpose and learn from its findings. Families have also been given photos of the children that are taken during data collection. These have proved very popular, as many families do not have access to cameras and have few photographs of their children. In Vietnam, the team evolved an innovative strategy in the early stages in which children from Children's Forums and Young Journalists Clubs, a nationwide programme run jointly by the Voice of Vietnam radio station and the Vietnam Youth Union, disseminated study findings directly to policy stakeholders in workshops and discussion fora. This enabled children to share Young Lives' findings and explore their implications for policy with local officials.

Key lessons from Young Lives about cohort maintenance

- Research reciprocity is vital to both cohort maintenance and data quality in longitudinal research, and researchers should develop clear strategies which guarantee that all exchanges with study participants are based on mutuality and respect. Programme managers should work together with principal investigators, field coordinators and supervisors to ensure a culture of respect for participants, their families and communities.
- Capacity building and retention of experienced field staff safeguard data quality and cohort maintenance in longitudinal research. Retaining fieldwork coordinators and field supervisors who are responsible for relationships of trust with respondents, as well as data collection logistics, training, tracking, piloting and cohort maintenance, is a priority.

Technological development in survey administration

During the first two rounds of household-based surveys, Young Lives used paper-based questionnaires. However, at Round 3 the study transitioned to computer-assisted personal interviewing (CAPI). The benefits of using digital technologies in research are widely acknowledged (Dunn and Banati 2015). The advantages of CAPI include the production of usable data in electronic form more quickly, easier data linking, and higher data accuracy due to its built-in data checks. Overall, the introduction of CAPI was beneficial for Young Lives due to its technical advantages, but there are risks. These include: cost, theft, technical glitches while in the field, translation, programming skip patterns, lack of familiarity with computers among enumerators, and the possibility of introducing bias or influencing responses.

The introduction of CAPI had significant implications for field teams. CAPI requires more work in advance compared to paper questionnaires, including programming skip patterns and pre-populating data to ensure that discrepancies between current and previous rounds can be flagged.

Concerns about the potential risks to survey administration and data management associated with the use of computers explains why CAPI was only introduced at Round 3. Practical considerations included the fact that in the early 2000s computer access was not particularly widespread in LMICs, especially in rural areas. This led to a concern that their use might disrupt the interaction between enumerators and respondents, possibly leading respondents to withdraw their consent to remain in the study. Other concerns included recharging computers while in the field, anxiety about the risk of loss or theft of computers, the possibility of technical glitches while in the field, and the cost and logistics of obtaining computers. Additionally, enumerators were accustomed to annotating paper questionnaires with their reflections on the interviews, enriching the research findings – a facility not available with software.

Is new technology always better? View from Peru on tracking

> “Administering a tracking questionnaire has always been an important part of our data management process to reduce attrition. When we transitioned to CAPI, Oxford was encouraging all the teams to do the tracking directly using tablets, but after many discussions we decided to use physical questionnaires. The enumerators felt that the paper questionnaires were easier for them, and quicker to administer. They preferred to write everything down on paper and enter the information into the tablets later on. This is one example of the way Oxford has been flexible. There is an awareness that some specific things can be adapted according to the context.”
>
> (Principal investigator, Peru)

Researchers were also concerned that introducing CAPI could affect respondents' attitudes towards the study, thereby introducing bias. However, a study by Escobal and Benites (2013) revealed that in most cases survey results were not affected by the introduction of CAPI. They noted the importance of the questionnaire design and training enumerators to ensure that their use of the technology did not impact their rapport with respondents and introduce bias.

The change to CAPI also highlighted the importance of ensuring that data-collection teams have diverse skills. Given that retaining experienced and outstanding interviewers has always been a priority, in Round 3 Young Lives invested heavily in training and developing the capacity of these enumerators to ensure they would be comfortable using both CAPI and the software.

Key lessons from Young Lives about the use of CAPI

- CAPI is a useful tool in survey administration and has distinct advantages over paper questionnaires, such as producing usable data in electronic form more quickly and built-in validation rules that help ensure data quality.
- It is vital for data and research teams to be aware of the risks of introducing CAPI technology into a study, and to exercise flexibility if possible.
- Training in the use of CAPI and software is essential for effective use of the technology.

Data management

Data management is a complex and technically demanding endeavour that encompasses data storage, transfer, access and use, while also ensuring data quality and security. Young Lives has developed data management practices to facilitate cross-cutting analyses based on the mixed-methods, multi-component, longitudinal dataset.

Directed by a central data manager in Oxford who matrix-manages the four study-country data managers, the data-management team works closely with principal investigators, research assistants and other researchers in ensuring that the data are well-organised, accessible, clean and of the highest quality. The team also plays a vital role in developing and implementing policies and procedures for safeguarding data security, particularly with regard to respondents' personal data, these being any data that enable the identification, whether directly or indirectly, of study participants. Data management involves an array of tasks, specifically verification of questionnaires, construction of databases with built-in validation rules, and editing, cleaning, confirmation and crosschecking of data, as well as checking for logical relationships within and across data forms and rounds. In addition, systems are in place to allow researchers to link the qualitative, school survey and household survey data and to track all the data for a particular individual across the different waves of interviews and research components.

The value of teamwork and sustained relationships: data managers' views

"I think for me, the people involved in the project are important. We must be able to work together. It's easier to work together when we know each other very well. If we change the staff, we have to take time to get to know them, and how to work with them, and they have to take to time to learn about the project."

(Data manager, Vietnam)

"The countries where we were able to keep the [same] data manager … were the easiest to work with. The data was the cleanest. It's important, if you can, to maintain consistent staff and invest in their capacity. We kept consistent staff, and made sure they were happy and trained, and were supported with everything they needed to get their job done well."

(Data manager, Oxford)

"Communication is key. We are always in communication with each other and know each other well after years of working together. For me what was very important was in 2012, after Round 3, the data managers in all the study countries went to Oxford for training about CAPI and it helped us get to know each other personally. We always stayed in touch after that. We helped each other with any challenges that came up, especially in terms of programming CAPI. We relied on each other. Of course, some people left, but many others stayed, and we have developed successful working relationships based on personal connections and communicating a lot."

(Data manager, Peru)

"We had excellent relationships with the data managers in all four countries, which was really helpful. Before, during and after data collection, we were in touch with them. I also met data managers in person when they came to Oxford for training sessions in CAPI. Each data manager had different skills and personalities, but they were all great to work with. The research assistants were also all really good and hardworking, and they drove the process! Close collaboration between research assistants and data management processes was really valuable. We felt that we were all working together on a valuable project. We were all on the same team. Not only did we work across countries, but we also worked across an interesting point in time in terms of technology. I remember my time at Young Lives fondly, as everyone was so nice, and we had such good relationships with everyone."

(Data coordinator, Oxford)

Key lessons from Young Lives about data management

- Conserving respondent anonymity and confidentiality is of paramount importance in longitudinal research and necessitates the development of a strict protocol safeguarding their personal data and location, along with secure information technology systems for data entry, storage, transfer and use. To be effective, such a protocol should define who controls and accesses which types of data (personal and anonymised) and to what ends, and specify the measures needed to prevent and address accidental loss, destruction, damage, or breaches.
- Programme managers must ensure that data managers have sufficient authority, support, and the technical skills needed to enable them to monitor and apply in full all data procedures within a study.
- Responsible for administering information technology systems and data protocols, and in some cases also for data archiving and data visualisation, data managers are vital to successful longitudinal research: they should be aware of and as far as possible involved in all data governance and data management decisions.
- Data cleaning, anonymising, storing, transfer, consistency checks and matching all contribute to ensuring data quality and security, and it is extremely important for data management and research teams to have the technical skills to oversee these tasks. They should have regular opportunities to learn about and whenever possible, take advantage of, new developments in the field.
- Effective data management facilitates cross-cutting analyses based on mixed-methods, multicomponent, longitudinal datasets.
- Good quality and strongly empowered country partners are key to generating high-quality longitudinal data, and also help ensure the research achieves national and global policy impact.

Data democratisation and discoverability

Young Lives aims to democratise its survey data by making them open access and therefore available as a public good. The survey data are archived with the UK Data Service together with the questionnaires, detailed supporting documentation and a constructed panel dataset of key variables across the survey rounds, ensuring that the data are available to all researchers and aiding their analysis (Briones 2018). In addition, Young Lives has evolved a range of measures to increase discoverability of its survey data – raising awareness of the data and potential applications for policy as well as promoting their use, including in study countries and by non-expert users. These measures include highlighting accessible data visualisations on the Young Lives website, and workshops promoting the use of the survey data among researchers and policymakers working with children and young people in study countries. These workshops have trained hundreds of students, academics, policy stakeholders and other professionals in the use of Young Lives data. Training workshops in Peru have shown government staff how to access and use Young Lives data for their own analysis and, eventually, for decision-making. For example, research officers from the Ministries of Education and the Economy and Finance used these data to review the factors that contribute to youth unemployment. Following the release of data from new rounds, the Oxford team has held competitions for students and researchers to prepare papers based on these data, with several of these calls directed at students and researchers in the study countries. Young Lives has also collaborated with Oxfam to create teaching resources specifically designed to engage adolescent learners.

There are risks associated with making data open access, particularly when external researchers and collaborators seek to link Young Lives data with administrative and other datasets, a process that requires the use of personal data. While these requests are consistent with data democracy objectives, it is not possible to release participants' personal data as this would contravene the study's responsibility to safeguard respondent anonymity and confidentiality (Murtagh et al. 2018). Young Lives recognises the value of data linking for policy development and is keen to help accommodate such initiatives: to this end, it has devised specific protocols to prevent breaches of the security of personal data. The protocols detail the terms under which collaborators and designated external researchers can commission Young Lives staff to do the matching, and release the anonymised matched data to them for their use.

Qualitative data, on the other hand, cannot be accessed by external researchers and are not archived for public use. This is due to the time and attention needed to fully anonymise these data, which often contain visual images of respondents or information about their lives and circumstances that reveal their identities or location. In addition, stripping out the contextual information to protect respondents' identities can significantly impair data quality and the potential for data analysis, with the risk that researchers might draw erroneous conclusions (Morrow and Boddy 2014).

Key lessons from Young Lives about data democratisation and discoverability

- Data democratisation, or archiving anonymised data so that they can be made open access, is an important way of enhancing the profile of a longitudinal study and increasing its research outputs.
- Data linking is a valuable aid to policy planning and development but needs careful attention to data security to avoid breaches in access to personal data.
- Efforts focused on data discoverability and capacity building in panel data analysis can enhance the use and application of data by external researchers. If promoting data use by external researchers is a priority, it must be planned for and reflected in study deliverables and budgets.

References

Attawell, K. (2004) *International Longitudinal Research on Childhood Poverty: Practical Guidelines and Lessons Learned from Young Lives*, Young Lives Working Paper 11, Oxford: Young Lives.

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

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

Dunn, K., and P. Banati (2015) 'Strength in Numbers: How Longitudinal Research Can Support Child Development', UNICEF Office of Research–Innocenti, https://www.unicefirc.org/files/upload/documents/glori-report.pdf (accessed 15 September 2020).

Escobal, J., and S. Benites (2013) 'PDAs in Socio-economic Surveys: Instrument Bias, Surveyor Bias or Both?', *International Journal of Social Research Methodology* 16.1: 47-63.

Morrow, V., and J. Boddy (2014) *The Ethics of Secondary Data Analysis: Learning from the Experience of Sharing Qualitative Data from Young People and their Families in an International Study of Childhood Poverty*, NOVELLA Working Paper: Narrative Research in Action, London: Institute of Education, University of London.

Murtagh, M.J., M.T. Blell, O.W. Butters, L. Cowley, E.S. Dove, A. Goodman, and P.R. Burton (2018) 'Better Governance, Better Access: Practising Responsible Data Sharing in the METADAC Governance Infrastructure', *Human Genomics* 12.1: Article 24.

Young Lives is an international study of childhood poverty and transitions to adulthood, following the lives of 12,000 children in four countries (Ethiopia, India, Peru and Vietnam) since 2001. Young Lives is a collaborative research programme led by a team in the Department of International Development at the University of Oxford in association with research and policy partners in the four study counties. This report has been funded by the Economic and Social Research Council.

The views expressed in this report are those of the authors. They are not necessarily those of, or endorsed by Young Lives, University of Oxford, ESRC or other funders

UKRI Economic and Social Research Council

Young Lives, Oxford Department of International Development (ODID)
University of Oxford, 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