An integrated and collaborative approach to developing and scripting questionnaires for longitudinal cohort studies and surveys: experience in Life Study

Suzanne Walton UCL GOS Institute of Child Health, UK
Stelios Alexandrakis UCL GOS Institute of Child Health, UK
Nicholas Gilby Ipsos MORI, UK
Nicola Firman UCL GOS Institute of Child Health, UK
Gareth Williams Ipsos MORI, UK
Duncan Peskett Ipsos MORI, UK
Peter Elias University of Warwick, UK
Carol Dezateux UCL GOS Institute of Child Health, UK
c.dezateux@ucl.ac.uk

(Received December 2016 Revised July 2017) http://dx.doi.org/10.14301/llcs.v8i4.434

Abstract

Efficient development of questionnaires for longitudinal surveys and cohort studies as computer-assisted survey instruments usually entails close collaboration between scientific and fieldwork teams. We describe a system based on the use of a Structured Query Language (SQL) database established to maximise efficiency, minimise error and ensure clear communication of requirements across teams for 'Life Study', a UK-wide cohort study designed to recruit mothers, their babies, partners and non-resident fathers, with whom further contacts were planned at the outset. The use of the SQL database enabled construction and integration of different elements of the study, initially through creating a master copy of each variable. This supported swift and accurate creation of a range of outputs enabling, for example, review and approval of successive drafts and final specifications of questionnaires, efficient implementation of changes to variables, re-use of metadata specified at the outset, reduction of ambiguities for survey programmers, and efficient and accurate automation of questionnaire scripting. The SQL database was also used to generate the syntax to transform pilot data into formats specified for data archiving and for associated publication quality questionnaires. This innovative use of an SQL database for questionnaire development and scripting, and subsequent data processing and documentation, highlights the value of this approach in improving the quality and efficiency of longitudinal surveys.

Keywords

Questionnaire design, questionnaire programming, data processing, data collection, data documentation, metadata, longitudinal studies, cohort studies, surveys and questionnaires, Structured Query Language (SQL) database

Introduction

Large-scale birth cohorts and longitudinal surveys comprise a key data resource, enabling interdisciplinary and life course research. Innovations in study design and timing of contacts can add complexity to the task of designing and programming questionnaires and may require an integrated approach for multiple respondents and contacts, especially for family-based or household designs. The aim of this research note is to report our experience of using an SQL database to develop and script questionnaires, re-using metadata specified at the outset in the context of a complex interdisciplinary cohort study for those involved in the design and administration of similar longitudinal studies. We present here an overview of methodology and advantages, rather than a technical guide.

'Life Study' comprised an innovative design of two integrated samples – a 'Pregnancy Component', purposively sampled from a small number of areas, and a nationally representative 'Birth Component', with multiple respondents and contacts in each (Dezateux, Knowles, et al., 2016; Goldstein, Sera, Elias, & Dezateux, 2017) (Table 1).

Table 1. Summary of core 'Life Study' protocol: planned contacts with participants

Participant	Pregnancy Component			Birth Component	
	Pregnancy	six months	12 months	six months	12 months
Mother	Yes	Yes	Yes	Yes	Yes*
Resident Father / Partner	Yes	No	No	Yes	No
Non-resident Father / Partner	Some	No	No	No	No

• Telephone or web-based interview

Questionnaires were developed with input from clinicians, population, social and biomedical scientists, relevant stakeholders and experts, and in consultation with UK research and policy communities. The Life Study Scientific Steering Committee was responsible for the final selection of measures and instruments.

The core protocol required design of seven questionnaires with two further questionnaires developed as a result of two funded enhancements on the maternal microbiome and on non-resident fathers (Bailey et al., 2015; Ipsos MORI, 2016). The integrated design necessitated appreciable overlap in, and harmonisation of, the questions included so that data collected in both components could be presented to the user as a single dataset. In addition, the scientific specification and operationalisation of a number of surveys had to proceed iteratively and within very tight time frames.

The UK Data Archive describes a "research data lifecycle" (UK Data Archive, 2017) comprising creating, processing, analysing, preserving, accessing and re-using data. Similarly, Banks, Calderwood, Lynn and Angel (2009) described a data production line comprising scientific direction, study design, instrument design, instrument realisation, data collection, data processing and data documentation in a report from the Survey Resources Network aimed at identifying improvements to efficiency and quality in data collection, management and processing of longitudinal surveys. This report highlighted the desirability of an efficient data production line at each stage and avoidance of duplication of work, while noting that in most longitudinal surveys the different elements of the data production line are carried out by separate organisations. Widespread practice is to specify questionnaire instruments using word processing software, typically Microsoft Word. Banks et al. (2009) state that "this can lead to ambiguities, in particular with respect to complex routing or question structures. This approach also does not facilitate the capture of structured metadata that can then be used at other stages of the data production process." (p. 6). The authors highlighted the central role of effective capture and reuse of metadata in quality assurance, the associated efficiency gains and need for questionnaire specification tools which capture metadata in a way that is both minimally ambiguous and maximally re-usable by downstream processes. Recognising that manual transfer of metadata between applications is time-consuming and error prone, they concluded that "capture of this metadata in a machine-parseable form is thus an important aspect of gaining control over the survey process as a whole and increasing the efficiency and quality of it" (Banks et al. 2009, p. 15).

The Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences (DDI Alliance, 2017). While a number of DDI tools were available at the time 'Life Study' was being developed, they were not mature or straightforward to use, and after discussion with DDI developers none were considered to fully meet our requirements, especially given the complexities of developing long questionnaires with inevitable multiple revisions, for multiple sweeps and respondents simultaneously. We elected to create a database to manage development of the 'Life Study' surveys which could be used at various points in the data lifecycle to overcome the limitations of more conventional approaches to scripting highlighted by Banks et al. (Banks et al., 2009). This approach built on existing skills and expertise within the team – allowing questionnaires to be developed as rapidly and efficiently as possible – and because the metadata was collected in a structured format, this enabled future interoperability with DDI-compatible tools and metadata.

Development of the questionnaires

Requirements

Questionnaires were developed on a modular or topic basis (Table 2). Additional survey elements supported collection of consent, biosamples, and a range of measurements, assessments and observations. Each questionnaire was planned to take 30-55 minutes and included interviewer administered (Computer Assisted Personal Interviewing - CAPI) and self-complete questions (Computer Assisted Self Interviewing - CASI) with complex routing. As there were over 1,840 questions/variables in the dataset – many to be used more than once – we required a system whereby a single, master copy of each variable could be specified, updated and applied to each questionnaire, with the ability to export these in different formats.

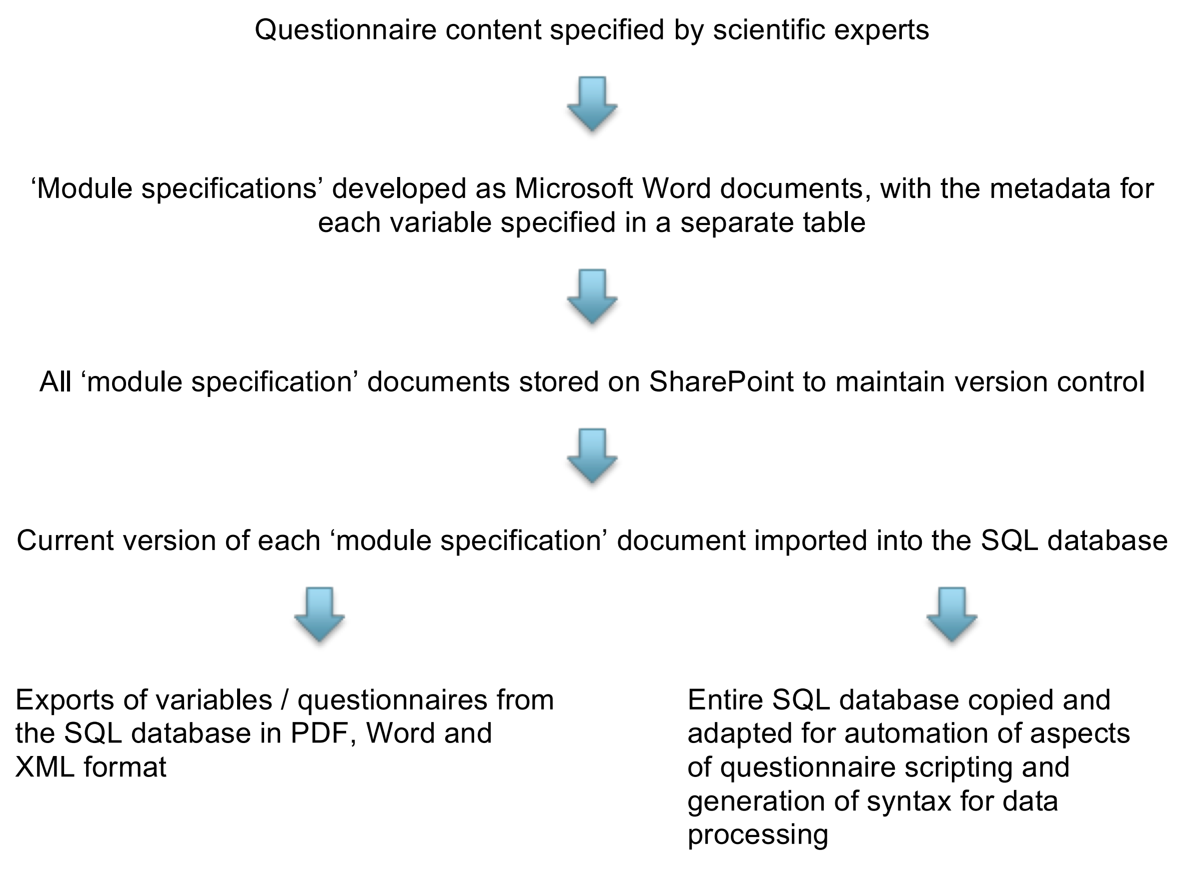
Table 2. Life study modules

1. Demographics
2. Identity
3. Parental and Family Health
4. Parental Mental Health
5. Parental Behaviour and Lifestyle
6. Parental Education
7. Parental Employment
8. Financial Situation
9. Pregnancy and Birth
10. Child Health
11. Child Development
12. Child Sleeping and Crying
13. Diet and Nutrition
14. Infections and Immunity
15. Childcare
16. Parenting
17. Family Relationships
18. Social Networks and Support
19. Housing
20. Neighbourhood
21. Environment
22. Partner Proxy
23. Consents
24. Interviewer Observations

Development of the questionnaire database

The 'Life Study' questionnaire database was developed as a client-server system comprising a database server and desktop client, using Microsoft SQL Server – a relational database using Structured Query Language (SQL). The customised application (client) was built using the Microsoft.NET framework to interact with the SQL database. The installation process entailed unzipping the client on a desktop computer with network/security permissions to connect to the SQL Server database. Questionnaires were developed and formatted so they could be imported into the SQL database (Figure 1).

Figure 1. Development of the Life Study questionnaire database



Questionnaire content specified by scientific experts

'Module specifications' developed as Microsoft Word documents, with the metadata for each variable specified in a separate table

All 'module specification' documents stored on SharePoint to maintain version control

Current version of each 'module specification' document imported into the SQL database

Exports of variables / questionnaires from the SQL database in PDF, Word and XML format

Entire SQL database copied and adapted for automation of aspects of questionnaire scripting and generation of syntax for data processing

Questionnaire specification and metadata

We specified questionnaire content in Microsoft Word for importing into an SQL database – as changes could be tracked, there were options for formatting, while spelling errors could be easily identified and corrected. We created one Microsoft Word document "module specification" for each module and, within these documents, developed a template/table containing the relevant information and metadata for each variable (See Appendix 1 - Explanation of the metadata for each variable). The tables were ordered in the sequence asked in the questionnaires. Successful import into the SQL database depended on maintaining – unaltered – the given formatting and headings of the tables within the Word document. The module specifications were populated with relevant metadata (Appendix 2 – Examples of variable tables). Fields within the table included: question name, variable label, question type (e.g. choice, date, number), and question text. The tabular structure ensured completion of all fields and a field indicating whether a question had been dropped enabled questions to be reinstated easily, if required.

Each table in the module specification documents had a field called 'universe'. These universe statements specified which questionnaires each variable was to be included in, and under what circumstances, and so were key to the functioning of the systems developed. The universe statements were based on Boolean logic and contained information on respondent types, study components, timings (sweeps) and modes of interview as well as, where relevant, conditioning based on responses to other variables. They were written in a standardised format so that the SQL database software could parse (automatically read and process) the information and export questionnaire documents and metadata accordingly. Specification of module order in conjunction with the order of tables/variables within a module and the universe statements stipulated the routing for each questionnaire.

Version control

A single master document for each module was stored on Microsoft SharePoint. This online document management platform enabled retention of all earlier versions and version control and reduced errors arising from multiple documents and multiple users.

Importing modules into the SQL database

The module specification documents were parsed as entire documents and imported into the Microsoft SQL Server database using the desktop client. The process of importing the modules was designed to identify errors in the universe statements or other problems arising from, for example, the merging of tables. Modules already in the database were deleted before importing a newer version. As module specifications were imported, a major version was created in SharePoint, date and time were automatically recorded and a comment added to indicate successful upload.

Database exports

The modules could be exported from the database in Microsoft Word and Excel formats, and also in Extensible Markup Language (XML). Export of modules was possible either individually – or as multiple modules simultaneously – using the desktop client in a number of pre-specified layouts based on component, respondent type, sweep and mode. This allowed the questionnaires to be displayed in a variety of formats for different audiences. These summarised variable name, question text, interviewer or interviewee instructions, response types and options, and the universe statement. A more detailed format was provided for computer programmers, called ‘programmer exports’, which included information needed to script computerised instruments, including whether to allow “don’t know” and “prefer not to answer” responses, as well as soft and hard checks.

‘Life Study’ staff had direct access to the Microsoft SQL Server Management Studio software, a database management user interface that was also located on the same server. This allowed them to copy the entire SQL database or individual tables and convert these into Microsoft Excel format.

Questionnaire scripting

Survey programmers at Ipsos MORI (the fieldwork partner for the ‘Birth Component’ pilot) were sent a copy of the entire SQL database plus its exports to enable them to computerise questionnaires using SPSS Dimensions software. They produced additional exports directly from the database to automate aspects of questionnaire programming. These allowed automatic extraction of variable names, question text, interviewer instructions, response options, and numeric ranges, thereby avoiding manual transfer of this information. Additionally, instructions such as “include timestamp here” were specified consistently throughout the module specifications, allowing the programmers to search for relevant terms and automate additional aspects of programming.

Once a computerised questionnaire had been created, Ipsos MORI checked and verified the programming using the ‘programmer exports’ provided by the ‘Life Study’ team. These documents contained all the relevant information that needed to be checked. An Excel log of any ambiguities and errors identified on checking was then passed to the ‘Life Study’ team, who enacted any changes required to the module specifications and sent an updated version of the SQL database back to Ipsos MORI.

Processing of raw datasets

On conclusion of the ‘Birth Component’ pilot, Ipsos MORI transformed raw data, held in SPSS format, using tables from the SQL database. This required addition of variable labels, “don’t know” and “prefer not to answer” codes, and variable name prefixes compliant with ‘Life Study’ variable naming specifications. Tables from the SQL database were used to create lists of variable names, labels and prefixes by module, which were copied into Microsoft Excel where formulae were used to generate SPSS syntax. This enabled SPSS syntax files comprising thousands of lines to be generated rapidly, and updated should errors be discovered. The addition of variable name prefixes and suffixes within the datasets specified the respondent and sweep to enable the provenance of variables to be understood, and – together with the structured approach to variable naming – enabled datasets to be easily re-shaped for longitudinal analyses.

Data documentation

In October 2015, when funding for 'Life Study' was withdrawn (Dezateux, Colson, Brocklehurst, & Elias, 2016), a 'metadata export' from the SQL database was created for the purposes of data documentation. This followed the basic structure of earlier database exports and contained additional information regarding the provenance of each question. From these, questionnaire metadata (Walton et al., 2016a, 2016b, 2016c, 2016d, 2016e, 2016f, 2016g) and documents for archiving 'Birth Component' pilot data with the UK Data Service (Dezateux, 2016) were rapidly produced.

Discussion

We report here our experience of developing an innovative system based on a SQL database that provided a robust approach to delivering the scientific design of a complex multipurpose cohort study, thus supporting high quality data collection, documentation and review standards and enabling iterative review of the questionnaires by scientific advisory and steering groups to review within the timetable of the study. Strengths of this approach include reduction of ambiguities and errors during questionnaire programming by specifying metadata at the outset as part of the instrument design stage, and time savings resulting from automation of several processes. This approach also facilitated data documentation in Microsoft Word and PDF format, database exports during data processing, and generation of syntax to transform SPSS datasets collected from pilot studies.

This approach is comparable in certain respects to the relational database management system described by Olsen (Olsen, 2012) who, since the late 1980s, has been developing methods to allow for integration of software and hardware solutions to survey and questionnaire design, including approaches to data collection and management in longitudinal surveys. Olsen (2012) demonstrated the effective use of relational database management systems in achieving integration between different stages of a longitudinal survey while avoiding separate questionnaire scripting. To our knowledge, other UK cohort studies have not employed this approach.

We developed an innovative tool based on SQL for capturing metadata in a machine-parseable format for re-use at each stage along the data production line. This is a dynamic system that requires programming expertise to create, but no prior experience of SQL to use. We employed a systems architect with experience of C#, Microsoft .NET framework and SQL to carry out the initial programming work. The 'Life Study' team had no prior experience of computer programming or of working with a SQL database, and remained responsible for all day-to-day interactions with the database and communication with the programmers.

Errors were significantly reduced by having one master version for each variable – and the ability to apply it when relevant – as opposed to specifying each questionnaire separately. Any changes to the master document were implemented consistently each time that question was used. This significantly decreased the time required to enact changes. Ambiguities and instances where information was missing were also significantly reduced for the programming team, as questionnaires were developed in a very structured format and were highly specified.

Exports from the SQL database allowed the modules to be produced swiftly as a professional publication with a standardised format. Each module was created as a separate document – however, it was possible to create entire questionnaires by combining PDF or Word documents with modules in the correct order. Exports were created in a standardised and consistent format for all steering and scientific group meetings and this facilitated these discussions and decisions. These summary formats were also used for ethics applications and for sharing with scientific experts, fieldwork agency staff and interviewers, other 'Life Study' team members and collaborators, and, ultimately, for publication. In creating these metadata documents from the SQL database, the 'Life Study' team had complete control over their content, format and style. The advantages of this approach were that questionnaire metadata were available before a fieldwork partner was even appointed, and were not constrained by the software used to generate the computerised scripts by a specific fieldwork partner. This approach also makes it easier to change fieldwork partner during a longitudinal study (as scripting is more efficient and less labour intensive, thus reducing incumbency advantage) and therefore improves tendering competitions.

Survey programmers at Ipsos MORI initially considered working directly with the XML outputs, but, after discussions, chose to work with the SQL database directly. The SQL database structure was flexible enough to allow the programmers to tailor the export mechanisms to their requirements. After about a week of development time, they were able to produce exports from the SQL database that extracted variable names, question text, interviewer instructions and response options. This reduced questionnaire scripting time as well as typographical and other errors, compared to typical practice where all aspects of the computerised script are programmed manually. Additionally, the more laborious aspects of questionnaire programming were removed, freeing up programmer time to concentrate on more complex tasks such as programming of routing.

Use of the programmer exports to check the scripted questionnaires reduced errors in programming, as well as time taken to check and correct programming. Similarly, using the SQL database to create syntax to process the raw SPSS datasets made it much easier to identify and correct errors and significantly reduced the time needed for this task.

In addition to aspects already covered, Ipsos MORI experienced time savings in scripting as a result of, firstly, not having to create survey metadata – whereas conventionally this is created 'de novo' with each study. Secondly, having a consistent structure to all the modules made collating them into one script much easier and, thirdly, there were fewer amendments required after checking. Time savings were also made, in that it was easy to make changes to question texts so that scripting created for one component of the study could be re-used for the other component of the study. The time required in programming the export process from the SQL database was more than recovered and, had the study continued, additional time savings were anticipated at each stage.

We had planned to produce 'Life Study' metadata for researchers to interrogate in an interactive electronic format such as provided by UK Biobank (Biobank UK, 2016) and the Health Survey for England (UK Data Service, 2017), either by using the SQL database directly or via the XML exports. We also anticipate our methodology would enable rapid import of survey metadata into cross-cohort search engines (CLOSER, 2017). 'Life Study' was closed before these could be developed, however exports from the SQL database allowed rapid production of metadata in PDF format for final reporting and data archiving.

In summary, the approach described in this research note enabled a complex survey with multiple sweeps and multiple respondent types to be developed by multidisciplinary teams in an efficient manner. It allowed automation of aspects of computerised survey scripting, saving time and reducing errors, and enabled rapid production of a dataset and questionnaire metadata for researchers to use.

As far as we are aware, 'Life Study' is the first UK cohort study to use a SQL questionnaire database in this way, using metadata specified at the outset at various points along the data production line. While many studies will not have the same complexity as the initial sweeps of 'Life Study', we would encourage others embarking on longitudinal studies, especially those where variables are re-used between sweeps, to consider adopting this novel approach.

Acknowledgements

This work was supported by the Economic and Social Research Council [Grant numbers ES/J007501/1, ES/L002507/1, ES/L002353/1, ES/L012871/1, ES/N007549/1].

The 'Life Study' team would like to thank Daniel Wallis and Dr John Godfrey from Tessella Technology and Consulting (www.tessella.com), and Robert Ireland for their assistance with programming developments and modifications of the SQL database after the initial development and programming work.

We also acknowledge the assistance of NatCen Social Research, the fieldwork partner appointed to carry out the 'Life Study' 'Pregnancy Component' pilot, for assistance with initial development of the module specification templates.

References

Bailey, S., Townsend, C., Rodgers, A., Dent, H., Mallet, C., Tsaliki, E., . . . Field, N. (2015). 15. Acceptability of collection of multiple bio-samples to birth cohort participants: implications for large studies. In Abstracts of the UK Molecular Epidemiology Group Winter Meeting on Metabonomics in Molecular Epidemiology. Imperial College London, London, UK. November 28, 2014. *Mutagenesis, 30*(3), 459-466. https://doi.org/10.1093/mutage/gev012

Banks, R., Calderwood, L., Lynn, P., & Angel, G. (2009). *A Feasibility Study to Investigate Integrated Survey Data Collection, Fieldwork Management and Survey Data Processing Systems for Longitudinal Studies*. Retrieved from http://www.researchcatalogue.esrc.ac.uk/grants/RES-234-25-0003/outputs/read/76474946-fba7-4a57-bfa6-b05a2414f661

Biobank UK. (2016). UK Biobank - Data Showcase. Retrieved from http://biobank.ctsu.ox.ac.uk/crystal/

CLOSER. (2017). About CLOSER Discovery. Retrieved from http://www.closer.ac.uk/data-resources/closer-search-platform/

DDI Alliance. (2017). Document, Discover and Interoperate. Retrieved from https://www.ddialliance.org/

Dezateux, C. (2016). *Life Study: Birth Component Pilot Study Sample, 2015: Secure Access [data collection]. UK Data Service. SN: 8072*. http://dx.doi.org/10.5255/UKDA-SN-8072-1

Dezateux, C., Colson, D., Brocklehurst, P., & Elias, P. (2016). *Life after Life Study: Report of a Scientific Meeting held at The Royal College of Physicians 14th January 2016*. https://doi.org/10.14324/000.rp.1485681

Dezateux, C., Knowles, R., Brocklehurst, P., Elias, P., Burgess, S., Colson, D., . . . Walton, S. (2016). *Life Study Scientific Protocol*. https://doi.org/10.14324/000.rp.1485668

Goldstein, H., Sera, F., Elias, P., & Dezateux, C. (2017). Integrating area-based and national samples in birth cohort studies: the case of Life Study. *Longitudinal and Life Course Studies 8*(3). http://dx.doi.org/10.14301/llcs.v8i3.439

Ipsos MORI. (2016). *Life Study: Qualitative work with lone mothers: Exploring options for contacting non-resident fathers*. Retrieved from UCL Discovery, London, UK: http://discovery.ucl.ac.uk/1485696/

Olsen, R. (2012). Infrastructure for Survey Data Processing in Urban and Planning Studies. In Carlos Nunes Silva (Ed.), *Online Research Methods in Urban and Planning Studies: Design and Outcomes* (pp. 17-36). https://doi.org/10.4018/978-1-4666-0074-4.ch002

UK Data Archive. (2017). Create and manage data - research data lifecycle. Retrieved from http://www.data-archive.ac.uk/create-manage/life-cycle

UK Data Service. (2017). Health Survey for England 2014 Retrieved from http://nesstar.ukdataservice.ac.uk/webview/index.jsp?v=2&mode=documentation&submode=abstract&study=http://nesstar.ukdataservice.ac.uk:80/obj/fStudy/7919&top=yes

Walton, S., Dezateux, C., Foster, N., Brocklehurst, P., Burgess, S., Colson, D., . . . Vignoles, A. (2016a). *Life Study Birth Component: Mother questionnaire*. Retrieved from UCL Discovery, London, UK: http://discovery.ucl.ac.uk/1485694/

Walton, S., Dezateux, C., Foster, N., Brocklehurst, P., Burgess, S., Colson, D., . . . Vignoles, A. (2016b). *Life Study Birth Component: Non-resident Father questionnaire*. Retrieved from UCL Discovery, London, UK: http://discovery.ucl.ac.uk/1485692/

Walton, S., Dezateux, C., Foster, N., Brocklehurst, P., Burgess, S., Colson, D., . . . Vignoles, A. (2016c). *Life Study Birth Component: Partner questionnaire*. Retrieved from UCL Discovery, London, UK: http://discovery.ucl.ac.uk/1485693/

Walton, S., Dezateux, C., Foster, N., Brocklehurst, P., Burgess, S., Colson, D., . . . Vignoles, A. (2016d). *Life Study Pregnancy Component: 6 month visit, Mother questionnaire*. Retrieved from UCL Discovery, London, UK: http://discovery.ucl.ac.uk/1485677/

Walton, S., Dezateux, C., Foster, N., Brocklehurst, P., Burgess, S., Colson, D., . . . Vignoles, A. (2016e). *Life Study Pregnancy Component: 12 month visit Mother questionnaire*. Retrieved from UCL Discovery, London, UK: http://discovery.ucl.ac.uk/1485708/

Walton, S., Dezateux, C., Foster, N., Brocklehurst, P., Burgess, S., Colson, D., . . . Vignoles, A. (2016f). *Life Study Pregnancy Component: Mother pregnancy visit questionnaire*. Retrieved from UCL Discovery, London UK: http://discovery.ucl.ac.uk/1485674/

Walton, S., Dezateux, C., Foster, N., Brocklehurst, P., Burgess, S., Colson, D., . . . Vignoles, A. (2016g). *Life Study Pregnancy Component: Partner questionnaire*. Retrieved from UCL Discovery, London, UK: http://discovery.ucl.ac.uk/1485676/

Appendix 1 – Explanation of the metadata for each variable

Field	Explanation
Question-Name	A unique variable name, consistent with the Life Study variable naming document
Variable-Label	Variable label to appear in the Life Study dataset

Allow-DK/REF	Indicates whether responses of "Don't know" (DK) and "Prefer not to answer" (REF) were permitted
	Yes = allow both
	Allow refusal only = allow "Prefer not to answer" but not "Don't know"
	No = don't allow either

Type	Choice: multiple = respondent can choose multiple response options Choice: single = respondent can only choose only one response option Control: = administrative e.g. introduction text. Open: XX = open text with fields limited to XX number of characters Number: A, B, C..D = number field, where A is the number of decimal places allowed, B is the maximum number of characters that can be entered, C..D is the range (lower and upper values) permitted Date: dd/mm/yyyy = date with day, month and year Date: mm/yyyy = month and year Date: yyyy = year
Source	Where the question originated from e.g. name of cohort study where used previously
Text	Wording of the question text to appear on the screen ^[xxxxxx] = text fill
TextBC	Text for Birth Component where this differs from the Pregnancy Component.
Interviewer-Instruction	UPPER-CASE TEXT signifies further instruction or information for interviewers in the CAPI sections USE CARD XX = there is a showcard for the interviewer to refer to. XX is the reference name/number of that card
Programmer-Instruction	Text fill instruction, including where to get the text fill information from and in which sweeps to implement the text fill.

	Loops - some variables are to be asked multiple times. The maximum number of loops is specified. Exclusive codes - some 'Choice: multiple' questions will have a response option which is exclusive and cannot be selected alongside other options. Include timestamp here = variable where time stamp required
Showcard	Yes = Showcard exists if the question is administered as CAPI. If there is no showcard, this field will remain empty If the showcard is different from the Options (then what is to appear on the showcard is given in this field).
Options	Table with response options
Help	Text for a help screen
Universe	Universe statement for the Pregnancy Component. RespType = 1 or 2 1 = Mother 2 = Partner Sweep = 1, 2 or 3 1 = Pregnancy Component pregnancy visit 2 = Pregnancy Component 6 month visit 3 = Pregnancy Component 12 month visit ModeType = 1, 2, 3 or 4 1 = CAPI (Computer-assisted personal interviewing) 2 = CASI (Computer-assisted self interviewing) 3 = PAPI (Pen and paper interview) 4 = Pre-visit paper questionnaire <> Is used to mean is not equal to @ sweep x means information to be fed forward from another sweep
UniverseBC6	Universe Statement for the Birth Component, 6 month visit. RespType = 1, 2 or 4 1 = Mother 2 = Partner 4 = Non-resident Father Sweep = 4 4 = Birth Component 6 month visit ModeType = 1, 2 or 4 1 = CAPI (Computer-assisted personal interviewing) 2 = CASI (Computer-assisted self interviewing) 4 = Pre-visit or post-visit paper questionnaire
UniverseBC12	Universe Statement for the Birth Component, 12 month visit. RespType = 1

	1 = Mother Sweep = 5 5 = Birth Component 12 month visit ModeType = 4, 5 or 6 4 = Pre-visit or post-visit paper questionnaire 5 = CATI (Computer-assisted telephone interviewing) 6 = WASI (Web-assisted self interviewing) <> Is used to mean is not equal to @ sweep x means information to be fed forward from another sweep
Soft-Check	Asks the participant to review their answer, as it exceeds the expected range.
Hard-Check	Indicates that an answer is not feasible. The answer needs to be changed before proceeding to the next question.

Flags	Who: This variable name prefix denotes who the variable refers to:
	m = mother
	p = partner
	c = child
	s = sibling
	g = grandparent
	h = household
	n = not applicable
	What: This variable name prefix denotes the type of data the variable is collecting:
	o = original
	d = derived
	y = physical measurement
	c = observations and assessments
	p = proxy report
	q = pre-complete question
	b = biological sample
	a = administrative data
	Multiple: yes This flag appears for questions which need to be asked separately for each child of a multiple birth.

Drop	Yes = This variable has been dropped and is no longer included in any of the questionnaires

Appendix 2 – Examples of variable tables within the module specification templates

Question-Name BabyDoB
Variable-Label Date of birth of cohort baby
Allow-DK/REF No
Type Date: DD/MM/YYYY
Source Millennium Cohort Study (MCS) First Survey
Text And what is ^[Cohort baby name]'s date of birth?
Interviewer-Instruction
Programmer-Instruction Include timestamp here Loop for each cohort baby as given at given at MultPreg @ sweep 1 or if sweep = 4,5 then loop for each cohort baby as given at NumBaby Textfill ^[Cohort baby name] from BabyName
Showcard
Options
Help

Universe
IF (RespType = 1) // Mother
AND (Sweep = 2) // 6 months
AND (ModeType = 1) // CAPI

UniverseBC6

```
IF (RespType = 1) // Mother
AND (Sweep = 4) // 6 months
AND (ModeType = 1) // CAPI
```

Soft-Check

If NumBaby>1 and babies are born on different dates:

"INTERVIEWER. THIS BABY WAS BORN ON A DIFFERENT DAY TO THE PREVIOUS BABY, PLEASE CHECK IF THIS IS CORRECT."

Hard-Check

IF date of birth is after date of interview: 'INTERVIEWER: This date is in the future. Please change!'

If date of birth is before 01/07/2014: "Answer '^[date entered]' is not in range '01/07/2014 - 31/12/9999'."

Flags

Who: c
What: o

Question-Name

ActSing

Variable-Label

Activities - sing

Use

Allow-DK/REF

Allow refusal only

Type

Choice: single

Source

Adapted from ALSPAC - Children of the children of the 90's (COCO90's)

Text

How often do you do these activities with ^[Cohort baby name]…

… Sing to ^[him/her]?

Interviewer-Instruction

Programmer-Instruction

```
Textfill for cohort baby’s name and sex
```

Showcard

Options

1	Every day
2	Several times a week - 2 to 6 times
3	Once a week
4	Less than once a week
5	Not at all

Help

Universe

```
IF ((RespType = 1) // Mother
And (ModeType = 2) // CASI
And (Sweep = 2)) // 6 month

OR ((RespType = 1) // Mother
And (ModeType = 4) // pre-visit
And (Sweep = 3)) // 12 month
```

UniverseBC6

```
IF ((RespType = 1 or 2) // Mother or Partner
And (ModeType = 2) // CASI
And (Sweep = 4)) // 6 month

OR ((RespType = 4) // Non-resident partner
And (ModeType = 2) // CASI
And (Sweep = 4) // 6 month
And (FrqSeChd = 1..7)) // Sees cohort baby
```

UniverseBC12

```
IF (RespType = 1) // Mother
And (ModeType = 5 or 6) // CATI or WASI
And (Sweep = 5) // 12 month
```

Soft-Check

Hard-Check

Flags

```
Multiple: yes
Who: m
What: o
```