5 SURVEY CONTENT

5.1 Wave 1 deposited data

The Wave 1 LSYPE dataset was originally deposited in December 2006. Since this time, extensive work has been undertaken to enhance the data by updating the variable names and labels and cleaning any inconsistencies within the file.

Three data files have been deposited for Wave 1, based on information collected from the young person (YP), the main parent (MP) and the second parent (SP). The following three files have been deposited as cross-sectional data files:

- Wave One: LSYPE Family Background file – May 2008
- Wave One: LSYPE Parental Attitudes file – May 2008
- Wave One: LSYPE Young Person file – May 2008

A fourth file represents the Household Grid information collected at Wave 1:

- Wave One: LSYPE Household Grid – May 2008

The Household Grid file is as a hierarchical file therefore containing one row for each individual identified in the household. This file contains a total of 70,643 cases, representing the 15,770 households who participated in the survey. The LSYPE Household Grid files are not deposited and are only available to approved researchers who make a request to the Department for Education (DfE) using the Confidentiality Agreement form available with LSYPE documentation on the UK Data Archive (http://www.esds.ac.uk/findingData/snDescription.asp?sn=5545#doc) or from team.longitudinal@education.gsi.gov.uk.

5.2 Wave 2 deposited data

Three data files have been deposited for Wave 2, based on the information collected from the young person, main parent and the second parent.

The following three files have been deposited as cross-sectional data files:

- Wave Two: LSYPE Family Background file – June 2008
- Wave Two: LSYPE Parental Attitudes file – June 2008
- Wave Two: LSYPE Young Person file – June 2008

A fourth file represents the Household Grid information collected at Wave 2:

- Wave Two: LSYPE Household Grid file – June 2008

This file is as a hierarchical file, therefore containing one row for each individual identified in the household. This file contains a total of 62,314 cases, representing the 13,539 households who participated in the survey. The LSYPE Household Grid files are not deposited and are only available to approved researchers who make a request to DfE using the Confidentiality Agreement form available with LSYPE documentation on the UK Data Archive (http://www.esds.ac.uk/findingData/snDescription.asp?sn=5545#doc) or from team.longitudinal@education.gsi.gov.uk.

5.3 Wave 3 deposited data

Three data files have been deposited for Wave 3, based on information collected from the young person, main parent and the second parent.

The following three files have been deposited as cross-sectional data files:

- Wave Three: LSYPE Family Background file – June 2008
- Wave Three: LSYPE Parental Attitudes file – June 2008
- Wave Three: LSYPE Young Person file – June 2008

A fourth file represents the Household Grid information collected at Wave 3:

- Wave Three: LSYPE Household Grid file – June 2008

This file is a hierarchical file therefore containing one row for each individual identified in the household. This file contains a total of 56,614 cases, representing the 12,439 households who participated in the survey. The LSYPE Household Grid files are not deposited and are only available to approved researchers who make a request to DfE using the Confidentiality Agreement form available with LSYPE documentation on the UK Data Archive (http://www.esds.ac.uk/findingData/snDescription.asp?sn=5545#doc) or from team.longitudinal@education.gsi.gov.uk.

5.4 Wave 4 deposited data

Three data files have been deposited for Wave 4, based on the information collected from the young person, main parent and the second parent.

The following three files have been deposited as cross-sectional data files:

- Wave Four: LSYPE Family Background file – September 2009
- Wave Four: LSYPE Parental Attitudes file – June 2009
- Wave Four: LSYPE Young Person file – September 2009

Two further files for Wave 4 are available but not deposited:

- Wave Four: LSYPE Household Grid file – June 2009
- Wave Four: LSYPE Activity History file

The Household Grid is a hierarchical file containing one row for each individual identified in the household. This file contains a total of 55,856 cases, representing the 11,801 households who participated in the survey. The Activity History file is also hierarchical, containing one row for each activity completed by the respondent since the prior interview. The LSYPE Household Grid and Activity History files are only available to approved researchers who make a request to DfE using the Confidentiality Agreement form available with LSYPE documentation on the UK Data Archive

(http://www.esds.ac.uk/findingData/snDescription.asp?sn=5545#doc) or from team.longitudinal@education.gsi.gov.uk.

Information on the activities completed by the young person respondent since the prior interview is alternatively available from the LSYPE Monthly Main Activity file for Waves 4 to 7. See Section 5.8 for more information on this file.

5.5 Wave 5 deposited data

Two data files have been deposited for Wave 5 based on the information collected from the young person.

The following files have been deposited as cross-sectional data files:

- Wave Five: LSYPE Family Background file – March 2010
- Wave Five: LSYPE Young Person file – March 2010

Two further files for Wave 5 are available but not deposited:

- Wave Five: LSYPE Household Grid file – March 2010
- Wave Five: LSYPE Activity History file

The Household Grid is deposited as a hierarchical file containing one row for each individual identified in the household. The file contains a total of 51,121 cases, representing the 10,430 households who participated in the survey. The Activity History file is also hierarchical, containing one row for each activity completed by the respondent since the prior interview. The LSYPE Household Grid and Activity History files are only available to approved researchers who make a request to DfE using the Confidentiality Agreement form available with LSYPE documentation on the UK Data Archive (http://www.esds.ac.uk/findingData/snDescription.asp?sn=5545#doc) or from team.longitudinal@education.gsi.gov.uk.

Information on the activities completed by the young person respondent since the prior interview is alternatively available from the LSYPE Monthly Main Activity file for Waves 4 to 7. See Section 5.8 for more information on this file.

5.6 Wave 6 deposited data

One data file has been deposited for Wave 6 based on the information collected from the young person. This has been deposited as a cross-sectional data file:

- Wave Six: LSYPE Young Person file – October 2010

Two further files for Wave 6 are available but not deposited:

- Wave Six: LSYPE Household Grid file – October 2010
- Wave Six: LSYPE Activity History file

The Household Grid is a hierarchical file containing one row for each individual identified in the household. The file contains a total of 49,838 cases, representing the 9,799 households who participated in the survey. The Activity History file is also hierarchical, containing one row for each activity completed by the respondent since the prior interview. The LSYPE Household Grid and Activity History files are only available to approved researchers who make a request to DfE using the Confidentiality Agreement form available with LSYPE documentation on the UK Data Archive (http://www.esds.ac.uk/findingData/snDescription.asp?sn=5545#doc) or from team.longitudinal@education.gsi.gov.uk.

Information on the activities completed by the young person respondent since the prior interview is alternatively available from the LSYPE Monthly Main Activity file for Waves 4 to 7. See Section 5.8 for more information on this file.

5.7 Wave 7 deposited data

One data file has been deposited for Wave 7 based on the information collected from the young person. This has been deposited as a cross-sectional data file:

- Wave Seven: LSYPE Young Person file – October 2011

Two further files for Wave 7 are available but not deposited:

- Wave Seven: LSYPE Household Grid file – October 2011
- Wave Seven: LSYPE Activity History file

The Household Grid is a hierarchical file containing one row for each individual identified in the household. The file contains a total of 45,839 cases, representing the 8,682 households who participated in the survey. The Activity History file is also hierarchical, containing one row for each activity completed by the respondent since the prior interview. The LSYPE Household Grid and Activity History files are only available to approved researchers who make a request to DfE using the Confidentiality Agreement form available with LSYPE documentation on the UK Data Archive (http://www.esds.ac.uk/findingData/snDescription.asp?sn=5545#doc) or from team.longitudinal@education.gsi.gov.uk.

Information on the activities completed by the young person respondent since the prior interview is alternatively available from the LSYPE Monthly Main Activity file for Waves 4 to 7. See Section 5.8 for more information on this file.

5.8 Monthly Main Activity deposited data

In autumn 2011, DfE deposited an additional data file relating to the activities of 11,821 young person respondents recorded at Waves 4, 5, 6 and 7. This information is considered to be the primary source of information for establishing a young person's 'main' activity at any point in time. Whilst it is also possible to obtain information on activities current at the time of interview from the Young Person file at each wave, this approach is not recommended for any analysis involving the comparison of activities due to the interview

period being spread across six months. At every wave these six months always crossed the start and end of two consecutive academic years and the traditional “summer vacation” period, which is known to often be a time of activity transition. Analysis of activities should therefore, where possible, focus on activities conducted in a single month as indicated by the Monthly Main Activity file.

The Monthly Main Activity file takes responses to the Activity History section of the questionnaire at each wave and synthesises this information into variables that represent a monthly time series running from September 2006 (two months after the respondents completed compulsory education) until May 2010 (the first month of interviews for Wave 7). The dataset has one row per respondent, making it easier to use than the Activity History files which are one row per activity.

Whilst the Activity History files provide up to fourteen different activity categories, information on the types of activities has been summarised into four categories for the Monthly Main Activity file. These categories closely match those activities of most interest to the policies of the Department for Education. The four activity categories provided are the following:

- Education
- Employment
- Apprenticeship/Training
- Unemployed/Inactive (NEET)

The activities listed are recorded in the file through 45 “*finact*” variables, which represent the 45 months of data that are available. Each “*finact*” variable takes a value that represents one of the four categories or states that there is ‘Insufficient information’. Where the latter occurs, this indicates respondents where the time series has ended prematurely. This is usually due to sample attrition or because the respondent was either not able to or refused to recall activities.

A small amount of editing of information from the Activity History files has been completed in the derivation of 'main' activities to improve the usefulness of the data and aid interpretation. For example, some activities which were missing start/end dates have had these dates imputed. Care has been taken to only impute dates in a window appropriate to the activities that precede and follow those where information is missing, also taking into account the dates on which the respondent was interviewed. In addition, where factual discrepancies occur between information collected in consecutive waves of the study, information collected at the interview closest to when the activity took place takes precedent. If the Activity History and Monthly Main Activity files are used in parallel, differences such as these will need to be taken into consideration.

Source data relating to the Activity History section of the questionnaire at Waves 4, 5, 6 and 7 are available from DfE using the Confidentiality Agreement form available with LSYPE documentation on the UK Data Archive (http://www.esds.ac.uk/findingData/snDescription.asp?sn=5545#doc) or from team.longitudinal@education.gsi.gov.uk.

5.9 How to link the datasets

All of the datasets have a unique serial number and each file can be linked on the variable - *surveyid*. This serial number is unique to the cohort member and therefore each family.

It is important that each file is sorted by *surveyid* in ascending order to link the datasets. A typical SPSS command to link files is shown in Box 1.

Box 1 Merging the datasets together.

```
GET FILE='C:\wave_one_lsype_young_person_file_16_05_08.sav'
Sort cases by surveyid (A).
SAVE OUTFILE='C:\wave_one_lsype_young_person_file_16_05_08.sav'

GET FILE='C:\wave_one_lsype_parental_attitudes_file_16_05_08.sav'
Sort cases by surveyid (A).
```

```
SAVE OUTFILE='C:\wave_one_lsype_parental_attitudes_file_16_05_08.sav'

GET FILE='C:\wave_one_lsype_young_person_file_16_05_08.sav'
MATCH FILES /FILE=*
 /FILE='C:\wave_one_lsype_parental_attitudes_file_16_05_08.sav'
/BY surveyid
SAVE OUTFILE='C:\wave_one_lsype_young_person_and_parental_attitudes_file.sav'.
EXECUTE.
```

5.10 Multicoded variables

Multicoded variables are obtained from questions where the interviewer is instructed to 'code all that apply'. Each response category has a separate variable in the dataset. For example, the main and second parents' education details have been stored within the datasets as multicoded variables, therefore if a main parent has answered that they are educated to degree level and have GCSE grades A*-C then they will have a 'yes' response in both of these separate variables.

5.11 Missing values

Due to the complexity of the information collected during the survey, a number of missing value categories have been adopted. These are shown in Box 2.

Box 2 Summary of missing values applied to the LSYPE data.

Valid Missing Values (included within published calculated percentages)

-1	Don't know – enables respondents to answer don't know to questions.

Invalid Missing Values (excluded from published calculated percentages)

-91	Not applicable – used to signify that a question did not apply to a respondent, usually due to routing.
-92	Refused – used to signify when a respondent has refused to answer a particular question.
-93	Question routing error – respondent asked question not relevant to their situation, despite being correctly routed.
-94	Insufficient Information – mainly used for derived variables and signifies that there is relevant information missing from source variables.
-95	Unable to classify / code – response cannot be allocated within defined code frame.

-96	MP/SP/YP unable to complete CASI section – used to signify that a respondent was unable to complete the self-completion section. This value label was also used to identify respondents who had used an interpreter.
-97	MP/SP/YP refused CASI section – used to signify that a respondent refused to answer the self-completion section.
-98	MP/SP not present – used to signify that a respondent was not identified for this part of the questionnaire module (i.e. respondent was a single parent).
-99	MP/SP/YP not interviewed – used to signify that a respondent was identified as eligible to answer the relevant questionnaire modules but was not interviewed (this may be due to a number of reasons, i.e. not being available on the day the interview was conducted).
-100	Respondent declined to answer sexual experience questions.
-996	No parent in household – used in later waves where a respondent may live away from parents.
-997	Script error - data missing for question
-998	Interviewer missed question – used to signify item non-response due to interviewer/CAPI error.
-999	Missing household data – used to signify cases missing some household level information from the respondent.

These missing values have been applied to the majority of derived variables where necessary, but some derived variables may have required additional missing categories. These are fully documented in the derived variable documentation for each wave.

5.12 Variable names

The LSYPE Wave 1 data originally used the variable names which directly corresponded with the questionnaire. However, since the original data was deposited the variable names have been renamed. This is to:

- enable users to clearly distinguish between the different waves of data for both cross-sectional and longitudinal analysis
- enable users to clearly distinguish between the different modules of the interview completed by the young person, the main parent or the second parent.

Each variable name in the data has been revised to include a prefix to identify the wave of the survey, followed by the variable name which directly relates to the questionnaire and then a suffix to identify who the question relates to, i.e.

MP, SP or YP. Multicoded variables will also include an alphabetical suffix (two characters) and this will always be found at the end of the variable name.

The only time this procedure is not followed is when it has been necessary to create derived variables on the dataset. An example of this is where a 'flag' variable is derived when outliers have been edited for inclusion in a derived variable. Normally the raw data is left unedited and the change is made during the derivation of new variables. For example, in order to derive particular income variables, such as gross annual salary, it was necessary to check outliers and clean the data based on the assumptions made during these checks. The flag variable is therefore provided for those interested in specifically looking at the edited data and compare with the unedited data. These variables do not use a wave prefix but start with the word 'flag'.

A typical variable name is made up of the following characters:

[Prefix1] [Question name] [Suffix1].

A multicoded variable will use the following characters:

[Prefix1] [Question name] [Suffix1] [Suffix2]

Prefix1	Indicates the wave - ; W1= wave 1; W2 = wave 2 etc.
Question name	is directly comparable with the questionnaire. It is easy to search for questions within your dataset as long as you use the relevant wave prefix in front of the question variable name.
Suffix1	Indicates who the question was asked of -; YP= the Young Person, MP= the Main Parent and SP = the Second Parent.
Suffix2	Indicates a multicoded variable which can range from 0a = answer 1 to (for example) aw = answer 49.[10]

[10] It is highly unlikely that a multicoded variable will use more than 70 categories but this suffix system would allow for approximately 700 categories as the two characters of the suffix enable the multicoded

5.13 Variable labels

The variable labels included on the dataset were initially derived from the CAPI program. These have been reviewed in an effort to ensure consistency across waves and to clearly identify who the question was asked of and who the question related to.

In order to enhance the variable labels, the labels now include prefixes within the label to indicate the following:

- who the question was asked of i.e. MP, SP, YP;
- whether the variable was a survey administration variable, and,
- whether the variable was a derived variable.

The variable labels now use the following list of prefixes to clearly identify the source of the question;

HH	Household Section Interview
MP	Main Parent Interview
SP	Second Parent Interview
SP/MP	Second Parent information asked of either MP or SP[11]
YP	Young Person Interview
HR	History Section
DV	Derived Variable – this clearly identifies that this is a derived variable.[12]
ADMIN	Administrative data – this identifies when the question relates to the interviewer, for example, coding whether the self-completion section was completed.

Using MP, SP or YP as a prefix clearly identifies that the question was directly asked to that person, for example:

---

variable to go from 0a through to zz (0 representing the first 26 categories of answers and z representing the 27th grouping of 26 categories).

[11] Used in Wave 4 to denote questions about the SP that could be asked of Main or Second parent. The variable W4sourceSP indicates who answered these questions.

[12] Full details of all derived variables are available in the 'LSYPE Derived Variable Documentation' which has been deposited on a Wave specific level.

MP: Age first left education. SP: Age first left education.

If a variable is asked of the main or second parent but relates to the young person, this is also clearly defined in the labels, for example:

MP: Why YP no longer lives with natural parents

At Wave 5 a * was included in variable labels to indicate that the question or response categories differ to those from previous waves, even though the variable name remains the same longitudinally. For example:

DV: *Employment status of mother

5.14 Data cleaning

Each wave of the LSYPE data has gone through an extensive process of checks to ensure the consistency and validity of the data. These are checks that investigate any outliers found within the data, ensure that the data has followed the routing used in the questionnaire, ensure that the correct person has answered the relevant questions and ensure that information is consistent between directly comparable variables.

During the process of checking the data it was necessary to edit some responses and to create missing value categories to identify particular issues such as item non-response. For example, the Household Grid collects the relationships of each household member to the young person. If the information collected suggested that the young person was a parent to another household member (but this other household member was older than the young person) then this information would be edited. Edits are only carried out if a relevant correction is easily identified (for example, if we know the household member is actually the parent of the young person, then we would amend the relationship to indicate this). If we were unable to identify a

correction using the data available, (for example, the relationship is unknown) then a system missing value is created.

A number of variables have been derived to enhance the data; details of these are available in accompanying derived variable documentation. During the process of deriving certain variables it became necessary to edit the data within the derivation, leaving the raw data unedited. This mainly affected income derivations at Wave 1. Collecting data on income is notoriously difficult, as respondents may refuse to answer these questions and in other cases there are some obvious instances of respondent or interviewer reporting error (e.g. reporting an income of £32 per annum instead of £32,000). Where possible we have corrected this information within the derived variable.

A slight amendment has also been made to the household NS-SEC variables derived at Waves 3, 4 and 5. In previous waves, the NS-SEC variables were derived using the respondent's current or most recent occupation details. The Wave 3, 4 and 5 interviews only collected the person's current employment detail[13], therefore the NS-SEC variables are slightly different and have been given an amended name to identify this (the variable name will include a prefix of 'c' to represent this change)[14]. It should also be noted that at Waves 6 and 7 the NS-SEC variables have been derived solely using the responses of the current details of the young person.

5.15 Datasets

For the purposes of archiving the data it was necessary to remove a number of variables. Some of these variables relate to introduction sections within the CAPI programme. These variables are asked of the interviewer and therefore are not necessary within the dataset, although the question remains in the questionnaire documentation.

[13] In Wave 5, the current occupation information for the parents in the household were obtained from the responses of the young person.
[14] If users are interested in looking specifically at the differences, please refer to the derived variable guides for all previous waves which include the syntax used to create these variables.

A number of other variables have been removed from the dataset, which might compromise the anonymity of the young person and their families. This relates to variables such as the exact date of birth (although age at interview and year of birth are available) and any answers to open-ended questions such as longstanding illnesses (these are available as categorical variables only). As LSYPE has progressed, both legislation and guidance relating to the disclosure of personal information have changed. This means that on some occasions variables that were previously suitable for disclosure early in the study were no longer able to be disclosed at later waves.

Applications for permission to use variables not deposited as part of the main dataset can be made to DfE using the Confidentiality Agreement form available with LSYPE documentation on the UK Data Archive (http://www.esds.ac.uk/findingData/snDescription.asp?sn=5545#doc) or from team.longitudinal@education.gsi.gov.uk. DfE will consider applications on a case by case basis.

5.15.1 LSYPE Family Background file

The content of the Family Background file is summarised in Table 3, indicating the level of information available at each wave.

Table 3 Summary of content of LSYPE Family Background file.

Questionnaire Section and Content	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
Household section					
Languages spoken in the home	●	●	●		
Main parent section					
Family activities	●				
Household responsibilities	●	●	●		
Household resources	●	●	●	●	
Individual parent section					
Demographics	●	●		●*	
Qualifications and education	●	●		○	
Current activity	●	●	●	●	
Second adult current activity			●*	●	
Health	●	●		●	
Employment/activity history	●	●	●		
Employment training and earnings		●		●	
Benefits and tax credits		●			
Income estimates			●	●*	
Young person section					
Parental employment					●

● Section included
○ Section asked of boost respondents only
* Asked of MP only
Note: At Wave 6 and Wave 7 there is no Family Background file.

5.15.2 LSYPE Parental Attitudes file

The content of the Parental Attitudes file is summarised in Table 4, indicating the level of information available at each wave.

Table 4 Summary of content of LSYPE Parental Attitudes file.

Questionnaire Section and Content	Wave 1	Wave 2	Wave 3	Wave 4
Main parent section				
Attitudes to the young person's school and involvement in education	●	●	●	○
Extra-curricular classes	●	●	●	●
Year 10 subject choices	●	●		
Special educational needs	●	●	●	○
Parental expectations and aspirations	●	●	●	○
School history		●		
Year 11 experiences				○
Post-16 plans				○
History section				
Choice of current school	●			

● Section included
○ Section asked of boost respondents only
Note: At Waves 5, 6 and 7 there is no Parental Attitudes file.

5.15.3 LSYPE Young Person file

The content of the Young Person file is summarised in Table 5, indicating the level of information available at each wave.

Table 5 Summary of content of LSYPE Young Person file.

Questionnaire Section and Content	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7
Main parent section							
Special educational needs	●	●	●	○			
Relationship with young person and contact with services	●	●	●	○			
Reasons for not living with natural parents	●	●	●				
Risk factors (absences, truancy, police contact, bullying)	●	●	●	●			
School History		●					
Household responsibilities	●	●	●				
Household resources	●	●	●	●			
Young person section							
Demographics	●	●	●	●	●	●	●
Health and disability		●	●	●		●	●
Attitudes to school/teachers	●	●	●	●	●		
Year 10 subject choices and reasons	●	●					
Rules and discipline	●						
Homework	●	●					
ICT	●	●					
Study Support	●	●	●				
Future plans and advice	●	●	●				
Information, advice and guidance					●	●	●
Relations with parents	●						
Risk factors (truancy, bullying, smoking, drugs)	●	●	●	●		●	●
Household responsibilities	●	●	●				
Childcare and caring responsibilities	●	●	●	●	●	●	●
Use of leisure time	●	●		●			
Subjects being studied		●	●				
Knowledge of and intentions towards Apprenticeships and related schemes		●	●	●		●	
Current activities				●	●	●	●
Qualifications being studied for				●	●	●	●
Attitudes to higher education				●	●	●	●
Attitudes to debt				●	●	●	
Attitudes to work	●					●	●
Higher education					●	●	●
Potential higher education students						●	●
Education Maintenance Allowance (EMA)			●	●	●		
Jobs and training				●	●	●	●
Apprenticeships						●	●
NEET				●	●	●	●
Volunteering							●
Voting					●		●
Care to learn					●	●	
Attitudes on local area					●		
Income and benefits				●	●	●	●

● Section included
○ Section asked of boost respondents only

5.15.4 LSYPE Activity History file and Monthly Main Activity file

Data relating to the Activity History of young people are collected through a loop of questions asked at Waves 4, 5, 6 and 7. These questions look to record every activity that occupies the majority of the young person respondent's time at any given point between September 2006 (two months after completion of compulsory education) and the Wave 7 interview. Typical information collected through the Activity History includes the following:

- Type of activity (categorised as one of fourteen categories)
- Start/end date of activity
- Whether courses were completed
- Reasons for activity transition
- Whether illness/disability influenced change in activity
- Why periods of employment came to an end
- Activities being completed whilst unemployed
- Whether training accompanies periods of employment (at Wave 7 only)

The responses to the questions have been used to create the Activity History file at each wave, which represents the activities that took place between interviews. All resulting Activity History files are therefore one line per activity, with respondents who engaged in multiple activities having multiple lines. Those who did not change activity between interviews will not be included in the Activity History file; their main activity information is picked up from the Young Person file using the variables relating to the activity at the time of interview.

Activity History files for Waves 4, 5, 6 and 7 have not been deposited, however an alternative Monthly Main Activity file is available. This file synthesises the Activity History information from all four waves and derives the 'main' activity for each month from September 2006 to May 2010. This activity is summarised to be either Education, Employment, Apprenticeship/Training or Unemployed/Inactive (NEET). Where information was missing in the Activity History file, the derivation for the 'main' activity

either suppresses information for the months affected or, where a start/end date is missing, randomly imputes a start/end date in the appropriate date window.

Where it is required to perform analysis based on the activity of the young person respondent, it is recommended that data from the Monthly Main Activity file is used. This is best practice as it removes inaccuracies that are introduced by using the "current" activity at the time of interview, as interviews were conducted over a six month period. This approach will allow existing published sources to be matched as closely as possible.

The Activity History files for Waves 4, 5, 6 and 7 are only available to approved researchers who make a request to DfE using the Confidentiality Agreement form available with LSYPE documentation on the UK Data Archive (http://www.esds.ac.uk/findingData/snDescription.asp?sn=5545#doc) or from team.longitudinal@education.gsi.gov.uk.

5.15.5 LSYPE Household Grid

The LSYPE Household Grid files are not deposited and are only available to approved researchers who make a request to DfE using the Confidentiality Agreement form available with LSYPE documentation on the UK Data Archive (http://www.esds.ac.uk/findingData/snDescription.asp?sn=5545#doc) or from team.longitudinal@education.gsi.gov.uk.

The Household Grid files contain two types of information: individual identifiers and identifying characteristics (e.g. the person number of each respondent; their sex and age) and cross-sectional variables collected about everyone in the household (e.g. relationships between household members). At Waves 5, 6 and 7 considerably less information was gathered in the Household Grid and the variable containing information about the person's relation to the young person (e.g. W5Relation) was less detailed than in previous waves. Table 6 indicates the information available from the Household Grid.

Table 6 Summary of content of LSYPE Household Grid.

Summary of content
Household member**
Sex of person**
Relationship to young person**
Marital status of person
Whether person living with someone in the household as a couple
Employment status of person
Ethnic group of person
Position of Main parent
Position of Second parent
Position of mother
Position of father
Position of History respondent*
Position of HHgrid respondent

*At Wave 4 the History respondent is the Main Parent

**Indicates that this information was asked at Waves 5, 6 and 7, all other information is not applicable or was not asked in the Household Grid in these waves.

The Household Grid contains one record for each person who has ever appeared in the household for each family that participated at Wave 1[15]. The individual details pertaining to the young person, the main parent and the second parent (such as age, sex, marital status and family composition) are available on the cross-sectional files. Where applicable, these variables have been derived from the Household Grid and are included in the Family Background file and the Parental Attitudes file.

Any analyses relating either to the demographical information of other household members or to relationships between other household members, must be done using the LSYPE Household Grid files.

The LSYPE Household Grid file also includes individual level identifiers for the young person, main parent and the second parent that identify their position within the Household Grid and other datasets. The young person is fixed as person one in all households and can be identified by selecting on the variable

[15] At Waves 5, 6 and 7 it is possible that a small number of households will have members duplicated in the grid. This happened where the young person had moved out of a family home at W4 and then moved back into the previous household at Wave 5, or where a person moved out prior to Wave 5 and moved back in. Data collection methods did not allow any cross referencing with people who were not listed as being present in the household at Wave 4, therefore in these cases the household members were recorded as being new to the household at Wave 5. As there is less identifying information (such as date of birth, ethnicity) for household members at Wave 5, no assumptions have been made regarding previous household positions. The variable W5shgint ensures that a household member will

W1HHID=1 (or W2HHID=1; W3HHID=1; etc. depending on LSYPE wave). In Waves 1, 2, 3 and 4, both the main and second parent can take any position within the Household Grid. The main parent can be identified using the variable **W1MAINRES** (or W2MAINRES, W3MAINRES, W4MAINRES, depending on LSYPE wave). The second parent can be identified using the variable **W1SECORES** (or W2SECORES, W3SECORES, W4SECORES depending on LSYPE wave). From Wave 5 onwards main and second parents are no longer identified so these variables are not present.

A value of -98 used on the Secores variables in the appropriate wave indicates that the second parent is not present in the household. However, the Household Grid may still contain some details about these people (such as their relationship to the young person) even if they are no longer living with the young person. These parents can be identified by selecting on the variable **W2SHGINT=2** (or W3SHGINT=2, or W4SHGINT=2, depending on LSYPE wave).

5.15.6 LSYPE History file – Waves 1 and 2

The History file has been constructed using information collected at Waves 1 and 2 Most of the information was collected at Wave 1, but where the Wave 1 history interview was not completed, interviewers attempted to collect most of the information at Wave 2. Where the Wave 1 history interview was completed, the Wave 2 history interview was very short, covering only the relationship history. The content of the history file is summarised in Table 7. This file also includes a number of derived variables which are all outlined in the derived variable documentation. These derived variables link respondent information longitudinally.

---

not appear as being present in the household twice, even if the same person appears twice in the file. The variable W5NoUse can be used to identify cases where this appears to be a problem.

Table 7 Summary of content of LSYPE History file.

Respondent	Summary of content
Main Parent	
	Employment/Activity History since LAST INTERVIEW Current Activity Employment/Activity History for NEW ENTRANTS and respondents not interviewed at Wave one
Second Parent	
	Employment/Activity History since LAST INTERVIEW Current Activity Employment/Activity History for NEW ENTRANTS and respondents not interviewed at Wave one
History Section	
	Birth and Health Relationship History

6 WEIGHTING

This section explains the development of weights for LSYPE data, which were created to ensure any resulting analysis can account for the survey design for each wave. Section 6.1 discusses the preliminary process of deriving the Wave 1 weights. This weighting procedure was twofold, with pupils from maintained schools and those from non-maintained schools weighted separately. Weights for subsequent waves are discussed in sections 6.2 to 6.7. The correct method for establishing which weight to use for analysis is identified in Section 6.8 and details of how to specify the sample design using SPSS are provided in Section 6.9.

6.1 Wave 1 weights

In the first instance, a design weight was provided by the fieldwork consortium (this variable is called *designweight*). The value of this is the reciprocal of the pupil's selection probability scaled so that the weighted and unweighted achieved sample sizes were equal.

6.1.1 Weighting independent school pupils

Fifty-four schools from outside the maintained sector were sampled and 28 of these took part in the study. These 28 schools yielded 530 responses.

There was some variability in response probability depending on the sex of the student and the type of school (single-sex or mixed). To account for these differences, logistic regression models were used to establish which variables could be used to weight the data. From this, the only useful variables were the students' sex and their type of school. Cell weighting was then used to derive pupil non-response weights. The weights obtained are shown in Table 8.

Table 8 Response by type of school and sex of young person.

Category	No. responders	Weight
Boys in boys' schools	120	0.95
Boys in mixed schools	156	1.02
Girls in girls' schools	161	1.08
Girls in mixed schools	92	0.89

One respondent's sex was unknown, so they were given an average weight.

The initial design weights were trimmed and combined with the pupil non-response weights. Calibration weights were finally applied so that the achieved sample size matched the population breakdown by type of school (single-sex or mixed) and by region (London/not London). The population figures are given in Tables 9 and 10.

Table 9 Population proportions by type of school.

Category	Population proportion
Boys in boys schools	19.1%
Boys in mixed schools	31.9%
Girls in girls schools	27.7%
Girls in mixed schools	21.3%

Table 10 Population proportions by region

Category	Population proportion
London	18.9%
Rest of England	81.1%

6.1.2 Weighting maintained school pupils

838 schools were selected in the maintained sector and 646 of them took part.

Weighting the maintained sample consisted of three steps. First weights were calculated for school non-response, then pupil non-response was modelled within responding schools; finally calibration weights were calculated.

School non-response

Schools were measured on various explanatory variables. These included the proportion of pupils from non-White ethnic groups, the proportion with 5 or more GCSEs at grades A* to C, the deprivation status of the school (a binary

variable based on the proportion of pupils entitled to free school meals), and regional information. Logistic regression models were fitted for school non-response using all of these variables. The statistically significant terms were the school's deprivation status and its region. The school non-response weights were then calculated by cell weighting. These are shown in Table 11.

Table 11 Response by deprivation status of school and region.

Category	No. responders	Weight
Deprived stratum in London	57	1.23
Deprived stratum outside London	77	1.04
Not deprived stratum in London	70	1.07
Not deprived stratum outside London	442	0.95

Pupil non-response

The responding schools yielded 20,447 students. Most of these (about 97%) could be matched to National Pupil Database records, and thus there was good information on (for example) their gender, ethnicity, GCSE performance, Government Office Region (GOR) and free school meals entitlement. There was very little useful information on those without a National Pupil Database match (3%).

Those with a match to the National Pupil Database were processed separately to those without. Those without a match were given a mean weight, whilst for those with a match, logistic regression models were used to estimate the probabilities of response. The final logistic regression model included the terms GOR, ethnicity[16], qualifications[17], and an interaction term between GOR and White ethnic group. The weight was then calculated as the reciprocal of the response probability.

Calibration

Design weights were next combined with school non-response and pupil non-response weights to calculate combined weights which were calibrated to the

[16] Nine categories: White, Bangladeshi, Pakistani, Indian, African, Caribbean, Mixed, Other, Not Obtained/Refused.
[17] Three categories: Achieved Level 2 (5 GCSEs or equivalent at A* to C); Achieved Level 1 (5 GCSEs or equivalent at A* to G) but not Level 2; did not achieve Level 1.

population proportions given in Table 12. These proportions are sourced from the National Pupil Database.

Table 12 Proportion of young people by demographical breakdown.

Category	Proportion
Ethnicity	
White	83.0%
Bangladesh	1.0%
Pakistan	2.3%
Indian	2.3%
African	1.8%
Caribbean	1.4%
Mixed	2.2%
Other	2.5%
Not obtained	3.5%
GOR	
North East	5.3%
North West	14.8%
Yorkshire and The Humber	10.6%
East Midlands	8.9%
West Midlands	11.4%
East of England	11.0%
London	12.7%
South East	15.5%
South West	9.8%
Qualifications	
Not achieved level 1	10.5%
Achieved level 1, but not 2	33.0%
Achieved level 2	56.5%
Sex	
Male	51.0%
Female	49.0%

With the exception of pupils from London, the data were calibrated to marginal totals rather than the cell totals. This means, for example, that the proportion of White respondents will be the same in the weighted sample and in the population at Wave 1. Similarly, the proportion of pupils in the North East will be the same in the weighted sample and in the population. Despite this, the proportion of pupils in individual cells (such as White respondents in the North East) might vary between the weighted sample and the population. London was treated differently. The ethnic breakdown of pupils in London is very different from that in other parts of the county, therefore because responses among ethnic minorities in London was quite high, it was possible to calibrate respondents in London to their cell totals.

6.1.3 Combining maintained and independent school weights

The final stage was to weight the sample so that the maintained/independent school split matched the population proportions (92.5% maintained, 7.5% non-maintained). This weight variable is called *W1FinWt.*

6.1.4 Effects of the weighting

The purpose of weighting is to eliminate bias in the estimates of population quantities. However, when the calculated weights are very variable the weighting process will increase the random error in the estimates, thus reducing their precision. The effect the weights have on precision can be measured by their efficiency, or by the design effect (essentially the reciprocal of the efficiency). Table 13 shows the design effect and its breakdown.

Table 13 Design Effect.

Stage of weighting	Design effect
Selection weighting (after trimming)	1.250
Final weighting	1.276

This shows the design has an efficiency of 1/1.276=78.4%. The interpretation is that a simple random sample 78% as large as the achieved sample would give equally accurate estimates of national quantities. This is mainly due to the selection weighting: selection weighting accounted for a design effect of 1.25, but the non-response weighting and grossing to match population proportions increased the design effect by only 1.276/1.250= 1.02.

6.2 Wave 2 weights

This section explains how the data was weighted to account for the non-response between Waves 1 and 2.

A design weight was provided by the fieldwork consortium (this variable is called *designweight* and is available in the Wave 1 dataset). This is the reciprocal of the pupil's selection probability scaled so that the weighted and unweighted achieved sample sizes were equal.

There were a total of 15,770 productive or partially productive interviews in Wave 1, of which 15,678 were issued at Wave 2. Some of these did not respond in Wave 2. It is likely that the characteristics of the non-responding pupils were different from those of responding pupils, which could lead to biases in estimates of population quantities. A statistical model was used to model the differences between those who responded and those who did not. This enabled the derivation of non-response weights to reduce bias.

Logistic regression models were used to estimate a pupil's response probability, and the non-response weights were then calculated as the reciprocal of this estimated response probability. The non-response weight was combined with the Wave 1 weight (*W1Finwt*) to provide the Wave 2 weight (*W2Finwt*).

Different models were used to estimate the response probabilities of independent and maintained school pupils. The sample sizes for the two different groups are shown in Table 14 and the models described in Sections 6.2.1 and 6.2.2.

Table 14 Wave 1 and 2 sample sizes

Category	Wave 1 responders	Wave 2 responders
Independent	530	456
Maintained	15240	11383
Total	**15770**	**13539**

6.2.1 Modelling response from maintained school pupils

The Wave 1 data set included 15,240 pupils from maintained schools. The vast majority of these (14,674) had provided both Young Person (YP) and Main Parent (MP) interviews. This information was used in non-response weighting. The other 566 Wave 1 respondents were only partially productive in Wave 1 (missing either a YP or MP interview). These provided far less useful data to use for non-response modelling, and because of this different models were used for full and partially productive pupils.

Fully productive respondents

A logistic regression model was used to estimate the response probabilities of each pupil. The data were weighted by the Wave 1 weight (scaled to equal the achieved sample size) before modelling. Various variables were used as potential explanatory variables. Some were variables obtained from the sample frame (such as ethnicity, Government Office Region (GOR), type of school, pupil's qualifications). Others were socio-economic variables obtained from the MP's answers to the Wave 1 questionnaire (such as the MP's single parent status, current working status, income support status etc). Additionally, some were other answers to the Wave 1 questionnaire (such as use of cannabis, language spoken at home, and the pupil's plans on their education after reaching the school leaving age).

A forward stepwise logistic regression procedure was used to model whether or not a pupil responded. Nine variables were identified as statistically significant (at the 10% level). These nine and another three (YP's sex, School's admission status, and School's deprivation status)[18] were included in the final logistic regression model. The twelve variables are summarised in Table 15.

[18] The final three were included because they had been used in the stratification when the original selection was taken

Table 15 Variables identified for logistic regression module

Origin of variable	Variable	Comments on Wave 2 response
Sampling Frame	GOR	
	Ethnicity	African pupils were least likely to respond
	YP's qualifications	Those with level 2 qualifications were most likely to respond
	YP's sex	
	School's admission status	
	School's derivation status	
YP Wave 1 response	YP's plans for education after the age of 16	Those planning on leaving education at the age of 16 were less likely to respond
	YP having a computer at home	Those with a home computer were more likely to respond
MP Wave 1 response	Whether a single parent	Those from single parent families were less likely to respond
	Current working status	
	Whether MP claimed JSA	Pupils whose MP claimed JSA were less likely to respond
	Whether MP has an A-level	Pupils whose MP had an A-level were more likely to respond

Partially productive respondents

It was harder to find a useful model for non-response of the partially productive Wave 1 respondents. This was because their Wave 1 interviews contained less useful information, and also because of the small size of the dataset.

A final logistic regression model included three explanatory variables: ethnicity (made into a binary variable 'White' or 'not White'), qualifications (whether or not they had obtained Level 2) and sex.

6.2.2 Modelling response from independent school pupils

The Wave 1 data set included 530 pupils from independent schools. A forward stepwise logistic regression procedure, including several variables as potential explanatory variables, was used to model whether or not a pupil responded. The variables included in the final logistic regression model were whether the school was in London, whether the MP had a pre-1975 O-level, the YP's and MP's attitude to school and school work, the YP's sex and type of school (boys, girls or mixed).

In general, a positive evaluation of school (measured by whether the YP strongly agreed with the statement “School work is worth doing” and whether the MP was very satisfied with the YP’s progress at school) was associated with a high probability of response. Those whose parents had a pre-1975 O-level also had a higher probability of response. Pupils from London schools had a lower response rate.

6.2.3 Creating the weights

The reciprocals of the estimated response probabilities gave the unscaled non-response weights. The top and bottom 1% were trimmed and then scaled to have a mean of 1. These non-response weights (called *DesignweightSCALED*) range from 0.90 to 1.53 (SD=0.11236). The Wave 2 weight *W2Finwt* was calculated by multiplying *W1Finwt* by *W1toW2NrWt* and scaling to ensure they had a mean equal to 1.

6.2.4 Effects of the non-response weighting

The purpose of the non-response weighting is to eliminate bias in the estimates of population quantities. However, when the calculated weights are very variable the weighting process will increase the random error in the estimates, thus reducing their precision. The effect the weights have on precision can be measured by their efficiency, or by the design effect (essentially the reciprocal of the efficiency), as shown in Table 16.

Table 16 Design effects

Stage of weighting on Wave 2 data	Design effect due to weights
Using the Wave 1 weight *W1Finwt*	1.263
Using the Wave 2 weight *W2Finwt*	1.278

A rough interpretation is that the design has an efficiency of 1/1.278=78.3% relative to an equal probability sample taken with the same amount of stratification and clustering. The effects of the non-response weighting can be summarised by comparing the final two rows of this table. The design effect obtained using *W2Finwt* is only slightly greater than that obtained using *W1Finwt* .This means that the non-response weighting is associated with a high level of efficiency. This is because the high response rate and the fact

that response rates were quite similar among the main sub-groups led to very little variability in the non-response weights.

6.3 Wave 3 weights

This section explains how the data was weighted to account for the non-response between Waves 2 and 3.

A design weight was provided by the fieldwork consortium (this variable is called *designweight* and is available in the Wave 1 dataset). This is the reciprocal of the pupil's selection probability scaled so that the weighted and unweighted achieved sample sizes were equal.

6.3.1 Achieved sample

There were a total of 13,539 productive or partially productive interviews in Wave 2 (456 from independent schools and 13,083 from maintained schools). These resulted in 12,439 productive Wave 3 interviews (428 from independent schools and 12,009 from maintained schools). In addition, this includes two pupils who took part in Wave 1 but were non-responders in Wave 2 and were re-interviewed.

Pupils' non-response weights were calculated as the reciprocal of their estimated response probabilities. Following the same method used in Wave 2, these response probabilities were calculated separately for independent and maintained school pupils.

6.3.2 Weighting maintained school pupils

The vast majority of Wave 2 responders provided both Young Person (YP) and Main Parent (MP) interviews. This provided a large amount of data to use in non-response weighting. There was considerably less information on the small number of partially productive Wave 2 respondents. As there was also a considerable difference in the response rates between those who were fully productive and partially productive (with the fully productive respondents

having a much higher response rate) the Wave 2 weighting method of using different models for the two groups was repeated.

Fully productive respondents – Maintained schools

A logistic regression model was used to estimate the probabilities of each individual to respond to the survey. Data were initially weighted by the Wave 2 weight, scaled to equal the achieved sample size. Various variables were used as potential explanatory variables. Some were variables obtained from the sample frame (ethnicity, GOR etc). Others were socio-economic variables obtained from the answers to the Wave 2 questionnaire (such as single parent status, current working status, income support status etc). Additionally, some were other answers to the Wave 2 questionnaire (including use of cannabis, pupil's plans to leave school). A forward stepwise regression procedure was used to determine which variables should be included in the final model, with the final statistically significant terms being:

- Whether YP ever tried Cannabis
- Use of a home computer
- Ethnicity
- Payment for extra tuition (non-school subjects)
- Free School Meals
- Whether the YP does homework
- Job Seekers' Allowance
- Qualifications
- Whether still at same school
- Frequency of the MP's communication with the YP's teachers
- Whether YP currently thought to have special educational needs
- Suspension
- Tenure

A logistic regression model containing the variables given above plus the additional variables admission status and deprivation status was used to estimate response probabilities.

Partially productive respondents – Maintained schools

A logistic regression model was used to estimate the probabilities of each individual to respond to the survey. Data were again weighted by the Wave 2 weight, scaled to equal the achieved sample size. Fewer variables were available for non-response modelling – either because they hadn't been collected in Wave 2 or because sample sizes were too small to allow their use.

The same logistic regression model used in the previous wave was used with White ethnic group; Level 2 educational achievement and sex included.

6.3.3 Weighting independent school pupils

As with the maintained pupils, a logistic regression model was used to estimate the probabilities of each individual to respond to the survey. Once again, data were weighted by the Wave 2 weight, scaled to equal the achieved sample size.

A forward regression approach identified whether the Main Parent has a pre-1975 O-level and a variable about the YP being happy at school as being statistically significant. In addition an interaction term between sex and type of school was forced into the model.

6.3.4 Combining maintained and non-maintained weights

The final stage was to create a file containing the estimated response probabilities. The reciprocals of these probabilities gave the un-scaled non-response weights.

The non-response weights were scaled to have a mean of 1. The weights had low variability, mainly because of the high response rate. The weights for independent school pupils were left untrimmed, but the top and bottom 0.5% of the maintained weights were trimmed and then re-scaled. The final non-response weights ranged from 0.93 to 1.43 (SD=0.079), with the percentiles shown in Table 17. These were combined with the Wave 2 weights to create

the Wave 3 weights. After scaling to have a mean of 1 (SD= 0.533) the final Wave 3 weight percentiles described in Table 17 were obtained.

Table 17 Percentiles

Percentile	Non-response weight	Final Wave 3 weight
MIN	0.93	0.12
.5	0.93	0.14
1	0.93	0.16
2.5	0.94	0.18
5	0.94	0.23
95	1.16	1.82
97.5	1.26	2.14
99	1.38	2.56
99.5	1.43	2.99
MAX	1.43	5.31

The non-response weighting does not create a large loss of effective sample size. This is partially as the high response rate leads to low variability in the weights, but also because the non-response weights are negatively correlated with the Wave 2 weights.

6.4 Wave 4 weights

There were several stages to the Wave 4 weighting. Weighting was conducted for the main survey, the boost survey and a combined main and boost survey.

Both the main and boost cohorts incorporated weights accounting for the probability of being sampled to take part and a weight accounting for non response. Finally, the main and combined files were each weighted to the population.

6.4.1 Main cohort Wave 4

The weights for this part of the sample incorporated weights which account for the probability of being in the sample at Wave 4, weights to account for non-response and finally a population weight which ensured the profile of the sample was consistent with the profile that would be expected in the survey population.

The weights used to represent the probability of being in the sample at Wave 4 are the final weights from Wave 3. There were also a small number of respondents who did not take part at Wave 3, but who did take part at Wave 1 or Wave 2[19]. These respondents were given their weight from the most recent wave they completed, and this was multiplied by the mean final weight from Wave 3.

The weight accounting for the probability of being in the sample at Wave 4 was initially applied. Non-response weights were then created using logistic regression to determine a probability of taking part, using variables consistent with those investigated in the preceding three waves.

The two weights were multiplied together to give an overall weight for the main cohort. This weight was applied and then the data was rim weighted to the profile shown in Table 18.

[19] These were respondents who had either moved and been relocated for the Wave 4 fieldwork, or requested to rejoin the survey.

Table 18 Target Proportions for weighting at Wave 4

Category	Proportion
Sex	
Male	50.59%
Female	49.41%
School type	
Maintained	93.57%
Independent	6.43%
GOR	
North East	5.34%
North West	14.86%
Yorkshire and The Humber	10.64%
East Midlands	8.91%
West Midlands	11.43%
East of England	11.04%
London	12.48%
South East	15.48%
South West	9.82%
Ethnicity	
White/other/Not known	88.86%
Bangladeshi	0.98%
Pakistani	2.35%
Indian	2.29%
African	1.82%
Caribbean	1.45%
Mixed	2.26%

6.4.2 Boost cohort Wave 4

For the boost cohort entering the study at Wave 4, the weight to account for the probability of being sampled to take part in the survey was the same as that assigned at the sampling stage at Wave 1.

The design weight was applied and the non-response weights calculated using CHAID to determine a probability of response, which was then inverted to give a non response weight. The variables used in the CHAID were those that were available on the administrative data.

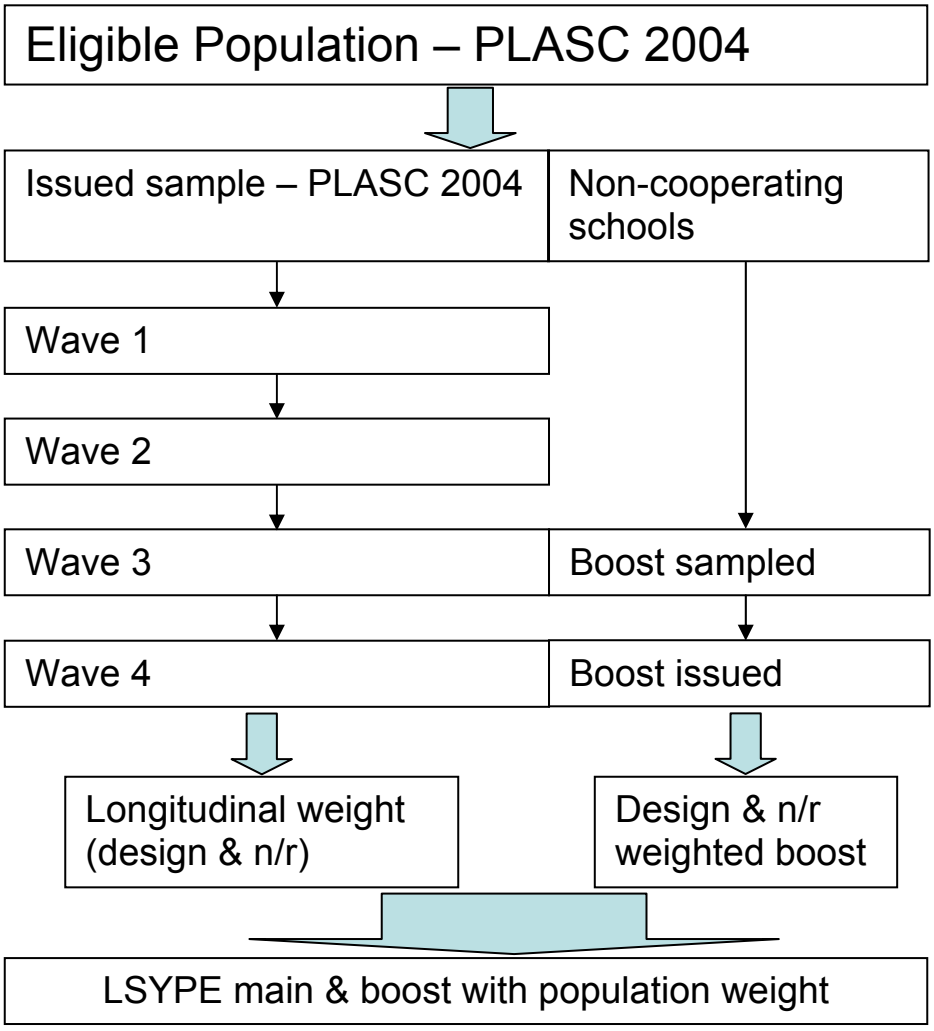
6.4.3 Combining main and boost

The initial design weights at Wave 1 were trimmed and scaled. This had to be accounted for when merging the main and boost files, so that the main and boost were each correctly represented in the combined file. Different factors

were applied to the main cohort and the boost cohort that adjusted for this based on the design weights at the initial sampling stage.

When combining the file, the design and non-response weights were applied to each of the main and boost files separately. Population weights were then applied so that the profile of the combined file matched the same population profile shown in Figure 1.

Figure 1 Combining main and boost weights



6.5 Wave 5 weights

Weights to account for non-response from certain groups between Waves 4 and 5 were calculated in two stages. Firstly the design weights were selected to account for the probability of being in the sample. At Wave 5, these were the final weights from Wave 4. With these weights applied, the profile of the

issued cases was then compared to that of the achieved cases, with regards to a range of variables from Wave 1. Similar to Wave 4, respondents from the main and boost cohorts had to be considered separately.

For the larger, main cohort, a logistic regression was carried out to see how well response could be predicted, however the models tested were poor predictors for non-response due to generally high response rates. The estimates generated by the model were not similar enough to the actual response rates generated among the subgroups, and so cell weighting was used instead.

A range of variables were tested, with those used for the non-response weights being combinations of sex and economic activity at Wave 4. The response rates for the groups that were used for weighting are shown in Table 19.

Table 19 Response rates used to calculate Wave 5 main cohort weights

Wave 4 Economic Activity	Male	Female
Full-time education	90.97%	93.31%
Full-time work	80.42%	84.85%
Part work & part training	86.75%	87.56%
Training course or Apprenticeship	89.01%	86.33%
Something else - Not NEET	82.81%	81.08%
Something else - NEET	69.77%	78.01%

For the boost cohort, there were too few people to consider logistic regression, therefore cell weighting was also used. The variable used for the cell weighting on this occasion was whether or not the sampled young person was studying for A Levels at Wave 4. The response rates within the groups are shown in Table 20.

Table 20 Response rates used to calculate Wave 5 boost cohort weights

Study at Wave 4	Response rate
Not studying A Levels	56.21%
Studying A Levels	72.86%

To obtain the final non-response weights, the inverse of the response rate was taken for each subgroup, and multiplied by the design weight to achieve

the final non-response weight. A final weight was then applied using rim weighting in order to match the profile of the respondents to that of the population. The target profiles are shown in Table 21.

Table 21 Target proportions for weighting at Wave 5

Categories	Proportion
Sex	
Male	50.59%
Female	49.41%
GOR	
North East	5.34%
North West	14.86%
Yorkshire and The Humber	10.64%
East Midlands	8.91%
West Midlands	11.43%
East of England	11.04%
London	12.48%
South East	15.48%
South West	9.82%
Ethnic Group	
White/other/Not known	88.86%
Bangladeshi	0.98%
Pakistani	2.35%
Indian	2.29%
African	1.82%
Caribbean	1.45%
Mixed	2.26%

6.6 Wave 6 weights

Weights to account for non response from certain groups between Waves 5 and 6 were calculated in two stages. Firstly, the design weights were selected to account for the probability of being in the sample. At Wave 6, these were the final weights from Wave 5. With these weights applied, the profile of the issued cases was then compared to that of the achieved cases, with regards to a range of variables from earlier waves.

There were two sections to the weighting. The first part was among those who took part in the study at Wave 5, and the other was among those who did not (Wave 5 skippers).

For those who took part in the previous wave, CHAID was used to identify groups who had statistically significant differing response rates. These included combinations of:

- Number of different contact details provided at previous wave
- Whether given permission to link answers with Department for Work and Pensions data
- Month of interview in previous wave
- Tenure in previous wave
- Mode of completion in previous wave
- Whether going to school or college at previous wave and level of qualification being studied
- How well teachers expected them to do in their Year 11 or earlier exams
- Whether applied for higher education in previous year
- Likelihood of voting in next general election

For those who skipped the Wave 5 survey, logistic regression looked at response propensities, and those that were found to be significantly related to response likelihood were:

- Number of different contact details provided by respondent at Wave 4
- Whether still living at same address in Wave 3 and Wave 4
- How likely they were to apply to university to do a degree in Wave 4.

Their propensity to respond was inverted to give a non response weight, which was multiplied by their final weight from the previous wave in which they took part.

The final weighted data for Wave 6 was compared to the weighted profiles for the previous waves to ensure that the key demographic variables were still in line with previous surveys.

6.7 Wave 7 weights

As with the previous wave, Wave 7 weighting involved 2 stages. Firstly, the design weights were applied, which accounted for the probability of being in the sample. For this wave, the final weights from the Wave 6 were used as the design weights. With these weights applied, the profile of the issued cases was compared with that of the achieved cases, with regards to a range of variables from the previous wave.

A logistic regression was carried out to estimate the probability of response among key groups that were associated with non-response or those considered to be of importance to be controlled for between the waves. These looked at the respondents' situation in the previous wave and how the sample compared on a range of measures. The final criteria that were controlled for were:

- Sex
- Ethnic group (with 'White', 'Other', 'Not known' combined)
- Tenure at Wave 6
- Survey mode at Wave 6 (internet, telephone or face-to-face)
- Interview month at Wave 6
- Higher Education application status at Wave 6
- Whether tried Cannabis by Wave 6 interview
- Whether had sex and at what age by Wave 6 interview

To obtain the final non response weights, the inverse of the probability of response was taken, and multiplied by the design weight to achieve the final non response weight. This was then trimmed at the 2nd and 98th percentile to remove any extreme weights, and then scaled to the achieved sample size.

Checks were done by comparing various statements for respondents who took part in Wave 7 with the Wave 7 weight applied, against those who took part in Wave 6 with the Wave 6 weight applied and assessed for similarities. None of the estimates looked at were considered to be out of an acceptable range of similarity.

No population weights were applied at Wave 7, as it was no longer possible to identify the up to date characteristics of the eligible population.

6.8 Weights to use in analysis

Every LSYPE wave with deposited data has an accompanying weight which is appropriate for analysis contained within a single wave. Where caution must be applied is where variables from multiple waves are used in a single piece of analysis. If this is the case, the general rule is always to use the weight from the most recent wave that a variable has been taken from. For example, if a cross-tabulation of attitude to school at Wave 1 by higher education application status at Wave 6 is completed then a weight from Wave 6 is required in order to complete robust analysis. This more recent weight is required to compensate for the demographic structure of the cohort changing over time as not all Wave 1 respondents remained in the study until Wave 6.

In some LSYPE datasets some specialist weights are additionally included to take account of particular situations. One example of this was the introduction of the sample boost at Wave 4, where in some cases additional weights are provided both with and without the boost cohort. Choice of weight in this situation should be made depending on whether variables from Waves 1, 2 or 3 are being used, as the boost cohort will not have variables from these three waves available. If such variables are being used then the boost cases should be excluded if the option is available.

There is an additional scenario with extra weights to cope with the introduction of young people who skipped particular waves of the study. As these “skippers” will not have responses to questions from the wave they missed, a weight with these “skippers” removed should be used if looking across waves. If analysis is only focused on a single wave then skippers can be included without any loss of accuracy.

6.9 Specifying the sample design

For more robust analysis (such as standard errors) it is preferential to specify the sample design. SPSS requires an additional module called ‘complex samples’ to specify the sample design. Box 3 provides the commands in both SPSS and STATA needed to specify the sample design.

Box 3 Syntax to specify the sample design

This example is based on Wave 1 variables .

SPSS COMMAND:

```
CSPLAN ANALYSIS
/PLAN FILE='enter file path and file name here.csaplan'
/PLANVARS ANALYSISWEIGHT=W1FinWt
/PRINT PLAN
/DESIGN STRATA= SampStratum CLUSTER= SampPSU
/ESTIMATOR TYPE=WR.
```

STATA COMMAND:

```
svyset [pweight=w1fintwt],psu(SampPSU) strata(SampStratum)
```

7 DATA LINKAGE

The LSYPE data has been linked to administrative data held on the National Pupil Database (NPD). The NPD is a pupil level database which matches pupil and school characteristic data to pupil level attainment.

We have also linked to school level indicators such as school size, proportion of pupils gaining 5 or more GCSEs at grades A*-C and truancy rates, and to geographical indicators such as the Index of Multiple Deprivation (IMD), the Income Deprivation Affecting Children Index (IDACI) and urban/rural indicators.

More detail on the three types of administrative data files currently linked to LSYPE can be found below. They can all be linked to the deposited survey data using *surveyid*.

National Pupil Database

The majority of pupils sampled from maintained schools have been linked to their NPD records. This file includes data on pupils' attainment at Key Stage 2, Key Stage 3 and Key Stage 4 and data about the pupil such as free school meal eligibility and Special Education Needs (SEN) status.

School Level Data

This contains information about the school each sample member attended at the sampling stage, and where we have linked to NPD, information about the primary school attended by the young person at Key Stage 2.

Geographical Data

Data from the National Statistics Postcode Directory (NSPD) have been linked by postcode. A small number of non-disclosive variables are included in the family background files for each wave.

Due to the potentially disclosive nature of some of these variables, the main linked administrative data described above have not been included on the

deposited LSYPE files. A reduced version has been deposited and includes information relating to number of school moves, free school meal eligibility, SEN and Key Stage 2 and 3 results.

Researchers requiring access to the fuller linked administrative files should contact DfE directly. Data are only available to approved researchers who make a request to DfE using the Confidentiality Agreement form available with LSYPE documentation on the UK Data Archive (http://www.esds.ac.uk/findingData/snDescription.asp?sn=5545#doc) or from team.longitudinal@education.gsi.gov.uk.

It should be noted that later waves of LSYPE have included consent questions to enable linking between LSYPE data and data held by the Department for Work and Pensions and the Department for Business Innovation and Skills. Whilst consent has been obtained for this additional linking and was already held for additional attainment information beyond Key Stage 4, as of November 2011 these data are not yet available.

8 APPENDIX A QUESTIONNAIRE AND DATA PROBLEMS

The following tables describe some of the problems identified during Waves 2 to 7 and also includes a column to identify what action was taken in relation to the data.

Table 22 Problems with Wave 2 questionnaire

Variable / section	Problem identified	Action taken
W2Modap3a	Question not asked if respondent did not say "yes" at Modap2.	Two variables have been created one for: 1) BMRB data only, and one for 2) All company data but filtered for Modap2 = "yes"
W2Disc1a	Question missing from the script	Variable in dataset only includes data collected by BMRB
W2NumAlev & W2NumGCSE	There was a problem in the creation of the sample variable that meant that most respondents were not asked these questions.	Data from respondents who did pass the filter has been left in the dataset
W2Hrefper	In a small number of cases (c.25) the Hrefper questions were answered by both parents.	Data left in for both respondents.

Table 23 Problems with Wave 3 questionnaire

Variable / section	Problem identified	Action taken
W3Plann16YP	NOP and BMRB used different filters therefore small number of respondents from NOP issued sample did not answer this question.	These variables are left in the dataset and a missing value has been added to identify the problematic cases on the variable W3Plann16YP. Further to this a new variable has been derived (W3Plan16YP) that incorporates both W3Plann16YP and W3RetedYP.
W3RetedYP	NOP and BMRB used different filters therefore small number of respondents from NOP issued sample who should have answered the W3Plann16 variable actually answered this variable.	
W3Exclude	This variable was actually used for text fill purposes rather than to measure whether a young person had been excluded from school.	This variable has been removed from the datasets.
W3AwareEMA & W3ApplyEMA	2 cases had not answered w3ApplyEMA when they should have been asked.	Due to the routing of these questions it was apparent that these cases suggested they were not aware of EMA (in W3AwareEMA) therefore have been set to No in W3ApplyEMA

Table 24 Problems with Wave 4 questionnaire

Variable / section	Problem identified	Action taken
W4NEETStatYP0a to W4NEETStatYP0g; W4NEET12YP	There was a mistake in the filter for the 2 NEET questions in the Word questionnaire. The filter in the Word questionnaire was: *{IF NEET (MainAct2 = 1 OR 2 OR 4) AND (HARMCHECK <> 2) AND (ExamChk = 2 }* However the filter should have been: *{IF NEET (MainAct2 = 1 OR 2 OR 4) AND (HARMCHECK <> 1) AND (ExamChk = 2 }* The script used by BMRB/MORI used the correct filter throughout fieldwork but the NOP script used the filter from the Word questionnaire at the beginning of fieldwork. As a result of this there are 35 cases who are missing from NEETStat and NEET12 who should have been asked these questions.	The cases have been coded to -997 'Script error'
W4TrainingYP	The filter in the Word questionnaire for the Training question is wrong. The filter before Training in the Word was as follows: *(Mainact =2 OR Mainact2=3) & (jobcol =2 OR jobcol=-1) & examchk=2* The problem with this filter is that respondents who are in part time employment (mainact2 = 3) never get asked jobcol and therefore can never qualify for the filter. The BMRB/MORI script didn't follow the word questionnaire and instead used the following filter: *(Mainact2 = 3 & examchk=2) OR (Mainact =2 & (jobcol =2 OR jobcol=-1) & (examchk=2))* The NOP script did use the filter in the Word questionnaire and as a result 113 respondents were not asked the Training question.	None
W4HE1YP0a W4HE1YP0b	At some point during the post pilot drafting for YCS and Lsype W4 the word "don't" failed to be deleted as intended from the Lsype Word questionnaire. However, as the BMRB script ran off the YCS version, their CAPI programme matched the YCS Word questionnaire as intended. The NOP script matched the mistaken Lsype word questionnaire.	None. Users should note this when looking at either or both of the variables
W4HEDecnYP W4Benefits_1 - W4Benefits_18 W4Costs_1 - W4Costs_18 W4HeCon_1 to W4HeCon_24 W4Debtatt1 - W4Debtatt6	There are a small number of cases (max 3) who have given responses to these variables when routing checks indicate that they should not have done. Their information has been retained in the relevant variables.	None
Second parent questions	BMRB/MORI only - At the beginning of fieldwork the BMRB script was not identifying the presence of second parents in the household. This could not	None. Any missing data will be coded appropriately

Variable / section	Problem identified	Action taken
	be picked up at the testing stage as it was something that only affected the live script version not the practice version that is used for testing. Most of this information was retrieved through re-interviewing.	
W4KS4check1YP	At the beginning of fieldwork the NOP script was not pointing correctly at the sample column holding the information about the young person's attainment. Most of this information was retrieved through re-interviewing.	None
W4AlevuniYP	For NOP cases the filter for this question was incorrectly placed within a previous filter which meant that although criteria were sometimes correct a lot of respondents did not get the chance to meet those criteria.	A flag variable (W4AlevflagYP) has been created to indicate whether the YP should have been asked the question.
StemAtt to Fundstud	Some NOP cases miss these questions due to an early script error which was resolved on the later releases.	Coded as script error
EverDDA	NOP filter was placed within a previous filter.	Coded as script error
Exten2	NOP filter for this question was placed in the wrong position which meant that the question could not be asked.	Coded as script error

Table 25 Issues with Wave 5 data

Variable / section	Problem identified	Action taken
W5Shgint	BMRB/Gfk-NOP edited the responses for a small number of cases where the young person had refused to say whether a household member was present in the household. This mainly applies to the young person's parent. This has led to a number of discrepancies when comparing those who appear to have a parent present in the household with the cases who have responded to the 'Parent' section of the interview	The variable W5ShgintFlag indicates the cases whose responses have been edited.
Household Grid	A small number of young people completing the online survey were 'swerved' around the household section as it was felt that the interviewer notes about household members would not be understood by the respondent.	Coded to -999 in the relevant variables
W5gender and W5relation (Household Grid file)	Where a person is no longer living in the household there is inconsistency regarding whether their sex/gender has been retained in the grid or has been coded to 'not applicable'.	None
W5relation	At Wave 5 there are fewer categories for the person's relationship to the young person than in previous waves. This is because of the nature of the mixed modes data collection where it may have been sensitive to allow the respondent to see more detailed categories.	None

Variable / section	Problem identified	Action taken
W5anyconb W5anycon	In the questionnaire, this question referred specifically to financial problems that the young person might encounter if they decided to go to university. However, many of the backcoded 'other' responses refer to problems that are not related to financial issues. Conversely, many young people gave responses relating to financial problems when answering W5anycon, which asked about problems other than costs and finance.	None, retained responses as given by the young person. Anyone analysing these variables may want to consider using both sets of responses.
W5BenftsYP0a to W5BenftsYP0i	Backcoding of 'other' responses in the Benfts variables means that the numbers of people answering the follow-up benefits questions do not always match up.	None, it was not appropriate to code these cases to 'script error' since the young person had responded 'other', meaning that correct routing had been followed.
w5ifuni	128 cases who didn't give a valid response in Unisubb were then routed to answer W5IfUni. The majority were completing the web survey where it is possible to enter a non-valid response (such as entering a space) and then to be routed as if a valid response had been given.	A flag variable W5IfuniFlag has been created to distinguish those who answered W5IfUni having given a valid response from those who were incorrectly routed.
W5EdExSubYP0a to W5EdExSubYP0d W5CITYSubYP0a to W5CITYSubYP0d W5CITYSubYP0a to W5CITYSubYP0d	As above, there is some inconsistency where the young person has entered an invalid response but the routing has taken them through to the follow-up question.	None
W5mselfYP / W5mmanysYP W5fselfYP / W5fmanysYP	For self employed people W5mselfYP/W5mmanysYP and W5fselfYP/W5fmanysYP don't always correlate, but the data collection allowed this.	None
Parental employment questions	Users should note that all information relating to parental employment is provided by the young person at Wave 5	None
SIC and SOC coding	With the introduction of a self-completion element to the survey SIC and SOC coding became slightly more difficult. Interviewers are experienced in collecting the correct information that is needed for this type of coding and will probe where necessary, therefore responses to the web survey were more difficult to code than the others and there may end	None

Variable / section	Problem identified	Action taken
	up being a higher proportion of uncodable responses for this mode at Wave 5	
Variables with low numbers that have been combined with other categories and removed from the file		
W5AnyconYP0g	Responses of young people who mentioned mental health problems have been combined with W5AnyconYP0f 'YP: Potential problems at university - Health problems (including mental health)/disability'.	W5AnyconYP0g not in dataset
W5AnyconYP0m	Responses of young people who mentioned drug problems have been combined with W5AnyconYP0z 'YP: Potential problems at university – Other answers'.	W5AnyconYP0m not in dataset
W5NEETProbYP0f	Responses of young people who mentioned 'YP: Barriers to becoming EET - Have my own children/ pregnant' have been combined with W5NEETProbYP0a 'YP: Barriers to becoming EET - Caring responsibilities'.	W5NEETProbYP0f not in dataset
W5NEETProbYP0i	Responses of young people who mentioned 'YP: Barriers to becoming EET - Mental health problems' have been combined into W5NEETProbYP0h 'YP: Barriers to becoming EET - Disability/ health problems'.	W5NEETProbYP0i not in dataset
W5HEsub5YP	Young people who were coded to 'DV: Subject area of degree would like to study: Agriculture & related subjects' have been combined with W5HEsub4YP 'DV: Subject area of degree would like to study: Veterinary science'	W5HEsub5YP not in dataset

Table 26 Issues with Wave 6 data

Variable / section	Problem identified	Action taken
w6BenftsYP0b	Two young people who are not employed have responded to this question, although the question specifies that it is referring to those who receive it 'not as an unemployed person'. Their responses have been retained in the data.	None
W6SexAgeYP	A small number of young people responded with very low ages to this question about age that they had first consensual sex.	These responses have been coded into category 1 ‘Under 14’.
W6SexSafeOften	These multicoded variables include a third response ‘Not applicable (respondent defined)’. Therefore these variables have three categories (1=Yes, 2=No, 3=N/a respondent defined’) rather than the two that would normally be expected with binary multicoded variables.	None
W6Wrk12YP	This question has been asked of the young person and parents in previous waves. However, in Wave 6 there was no response option for self employed young people who don’t employ anyone else. It may be that they responded ‘Don’t know’ instead. In derived variables that include this variable, anyone who is coded ‘Don’t know’ is coded to 0 employees in the derived variable.	None
Household Grid	A small number of young people completing the online survey were ‘swerved’ around the household section as it was felt that the interviewer notes about household members would not be understood by the respondent.	Coded to -999 in the relevant variables

Table 27 Issues with Wave 7 data

Variable / section	Problem identified	Action taken
w7NEETEduYP w7NEETWrk2YP w7AppEnableYP w7AppBensYP w7NEETDifOYP w7BullyDiscriminationYP w7BullyConsiderYP w7NumChiYP w7NEETMainJ w7OwnChi2	Unspecified routing problem led some respondents to incorrectly miss these questions.	Responses coded -997 “Script error”.
w7QuaWageYP	A small number of responses to this question were missing from the final dataset.	Responses coded -998 “Missing data - Question not asked”.
w7AlcEverYP	A small number of respondents were not asked this question due to a problem with a variable from a prior wave which should have been considered in routing respondents to this question.	Responses coded -996 “Problem with feed forward variable”.

Variable / section	Problem identified	Action taken
W7SexSafeOftenYP W7SexSafeOftenOYP W7SexSafeOftenO2YP	These multicoded variables include a third response 'Not applicable (respondent defined)'. Therefore these variables have three categories (1=Yes, 2=No, 3=N/a respondent defined') rather than the two that would normally be expected with binary multicoded variables.	None
w7BenftsYP0e	Despite correct routing being present in the questionnaire, this question was incorrectly asked only of respondents who declared their children at this wave. This meant that respondents who mentioned their children at Wave 6 and had no further children to mention at Wave 7 were not routed to this question.	As Child Benefit is not means tested it is a fair assumption that all the young people who were claiming it at Wave 6 would still be claiming it at Wave 7. To enforce that assumption the variable w7BenftsYP0e was removed from the dataset and replaced by the derived variable w7Benfts0e.
w7UnEmBenYP w7IncSupYP w7SkDsBnYP w7FamilyYP w7CCTCYP	Some respondents missed the opportunity to declare benefits in w7BenftsYP and instead chose to mention them in the non-dataset question BenftsO. The responses to this question were back-coded into w7BenftsYP and should have been used in the routing to w7UnEmBenYP, w7IncSupYP, w7SkDsBnYP, w7FamilyYP and w7CCTCYP. Unfortunately this last routing did not correctly function and so a small number of cases missed these follow-up questions.	Responses coded -997 "Script error - Free text response not routed to supplementary question".

9 APPENDIX B UPDATES TO WAVE 1 AND WAVE 2

Tables 28 and 29 below highlight some variable name changes that have been made on the data since archiving. These changes form part of an overall update to Waves 1 and 2 which took place in May 2008.

Table 28 Updates to Wave 1 data

Variable / section	Problem identified	Action taken
W1incareHH	These variables have been renamed to ensure consistency across all waves	W1InCarHH
W1intypeHH		W1InTypHH
W1evercarMP0i		W1whencarMP
W1scomad2HS		W1scomadi2HS

Table 29 Updates to Wave 2 data

Variable / section	Problem identified	Action taken
W2OwgherMP	These variables have been renamed to ensure consistency across all waves	W2OwherMP
W2Ben4bMP		W2Ben4AQbMP
W2Mhelp_1MP		W2Mhelp1MP0a
W2scomad2HS		W2scomadi2HS
W2OutschYP		W2OutschnYP
W2YouBulYP0a		W2youbulnYP0a
W2YouBulYP0b		W2youbulnYP0b
W2YouBulYP0c		W2youbulnYP0c
W2YouBulYP0d		W2youbulnYP0d
W2YouBulYP0e		W2youbulnYP0e
W2SCwhopreYP0a		W2dwhopreYP0a
W2SCwhopreYP0b		W2dwhopreYP0b
W2SCwhopreYP0c		W2dwhopreYP0c
W2SCwhopreYP0d		W2dwhopreYP0d
W2SCwhopreYP0e		W2dwhopreYP0e
W2SCwhopreYP0f		W2dwhopreYP0f
W2SCwhopreYP0g		W2dwhopreYP0g
W2SCwhopreYP0h		W2dwhopreYP0h
W2carehrs2YP		W2carehr2YP

10 APPENDIX C INDEX FILE

The LSYPE Index file is not deposited in the archive and is only available to approved researchers who make a request to DfE using the Confidentiality Agreement form available with LSYPE documentation on the UK Data Archive (http://www.esds.ac.uk/findingData/snDescription.asp?sn=5545#doc) or from team.longitudinal@education.gsi.gov.uk.

The Index file is a longitudinal file containing information of all household members including the young person, collected at Waves 1 to 7. The sample consists of the sample member (known as the Young Person) who was present and interviewed at Wave 1 and any members of their household who were present in the household at Wave 1 or any subsequent wave. This file is deposited as a hierarchical file containing one row for each individual who has ever appeared in the household.

The Wave 1 data forms the basis of the file and has been updated with the information from the subsequent waves. Therefore if a household member moved out of the household at any wave their details are still held in this file. If a new member has entered the household, the Index file is updated to include their information. This file represents all the 15,770 young people who participated in the study at Wave 1 plus 352 Wave 4 boost cases, and also represents all members of the young person's household identified at any wave.

The Index file contains individual identifiers (such as the person number of the respondents at all waves) and fixed characteristics (such as age, sex and relationship to the young person). The variables included in this file are described in more detail in Section 10.2.

The individual details pertaining to the young person, the main parent and the second parent (such as marital status and family composition) are available on the cross-sectional files (Waves 1 to 4). These variables have been

derived from the Household Grids and are included on the family background files and the parental attitudes files corresponding to each wave.

10.1 How to use the Index file

The Index file includes individual level identifiers for the young person, main parent and the second parent (where applicable) which identify their position within the Household Grid. The young person is always fixed as person one in all households and can be identified by selecting on the variable ***HHID*=1**. Both the main and second parent can take any position within the index file but can be identified for each wave using the variables ***WxMAINRES*** (for the main parent) and ***WxSECORES*** (for the second parent). This only applies to Waves 1 to 4, since from Wave 5 main and second parents were no longer identified in the interview. Similarly, the history respondent can take any position within the Index file and can be identified for each wave using the variables ***WxHISTRES***.

Users who wish to create person specific datasets i.e. with mother only responses or with second parent only responses can use a combination of the variables discussed in Sections 8.2.16 to 8.2.19.

The six variables ***WxHHRESP*** indicate whether there is information for an individual household member at a particular wave.

Survey level information is also available relating to Wave 1 to 4, providing users with an overall indication of response for each respondent (i.e. main/second parent, young person and history respondent) These variables are discussed in more detail in Section 10.2.22 onwards and will help users interested in longitudinal analysis of the data.

10.2 Variable descriptions

The first two letters of each variable contained within this dataset indicates the wave of the survey the variable refers to (for example, W1 refers to Wave 1

information and W4 refers to Wave 4 information), with the exception of the variable *NW3SPINT* (described in detail in Section 10.2.13).

Missing values within the Index file follow the same definitions as those described in Box 2 in Section 5.11.

For ease of reference, the cross-sectional variables are described below and are referred to as W*x* within this user guide, where applicable[20]. Details of how the derived variables have been constructed are provided in Appendix D.

10.2.1 surveyID

This is a unique household level identifier that can be used to merge this data and other deposited LSYPE datasets to each other[21].

10.2.2 HHID

This variable has been created as a 'fixed' person number across all waves. This was necessary as some household members were found to have swapped positions within the Household Grid between waves, although the majority of household members (including the main and second parent respondents) remain in the same position across all waves. Households that had swapped positions have been amended on the cross-sectional files and all will correspond with this variable.

10.2.3 Age

At Wave 1 each household member was asked their age and this variable records their answers. This variable is fixed based on the Wave 1 data but has been updated were possible with information collected at subsequent waves if details were missing at Wave 1. Dates of birth were collected in some subsequent waves (collected for all members at Wave 2 and new household members at Waves 3 and 4) and age was calculated using a combination of the date of birth and the Wave 1 interview date.

[20] In the dataset 'x' should be replaced with the numerical value representing the wave of interest
[21] See Box 5 for examples of how to match the Index File to the available cross-sectional datasets

10.2.4 Hdobm and Hdoby[22]

These variables identify the month and year of birth of all young persons, collected at the Wave 1 interview. Details of a household members' date of birth were only collected from Wave 2 and updated at Wave 3 and Wave 4 for any new respondents. These variables will therefore not include information for household members (other than the young person) who only responded at Wave 1 or who joined the household after Wave 4.

10.2.5 Sex

This variable is fixed across waves and refers to the sex of the individual identified within the household. This variable is fixed even for respondents who had swapped positions within the household as the Index file uses the variable *HHID* to form the longitudinal structure. The variable uses the Wave 1 data as a starting point and is updated accordingly with any new household members identified in subsequent waves.

10.2.6 ReltoYP

This variable identifies the relationship of each household member to the Young Person at Wave 1 (or at the first wave in which they joined). Whilst it is possible for relationships to change over time (due to their subjective nature), this is not captured within the Index file. Any users interested in identifying changes in relationships should refer to the wave specific Household Grids available. At Waves 5, 6 and 7 less detailed information was collected about the household member's relationship to the Young Person, so anyone joining the household after Wave 4 is coded to -989 in this variable.

10.2.7 ReltoYP2

This variable contains similar information to ReltoYP but is based on the reduced coding scheme for this question at Waves 5, 6 and 7. For those who were in the household prior to Wave 4, ReltoYP has been recoded into the

[22] It is possible that there are slight discrepancies between reported age and the Hdobm and Hdoby variables. This will mainly be were the day of birth results in a household member being a year younger or older when compared to the date of interview.

```
-997 "Script Error"
-996 "No Parent in household"
-970 "Response in W1 only"
-971 "Response in W2 only"
-972 "Response in W3 only"
-973 "Response in W1 and W2 only"
-974 "Response in W1 and W2 and W3 only"
-91 'Not applicable - second parent not present in HH'
0 'Second Parent interview not completed'
1 'Second Parent interview completed by second parent'
2 'Second Parent interview completed by main parent or other adult'
3 'Second Parent interview completed by main parent with consultation from second parent'.
execute.
```

B. Variables indicating change in main/second/history parent role

Variables 7 to 9 in Box 6 identify whether the main/second/history parent person number changed at any point across the first three waves. These variables identify whether the respondent was the same person at each wave or whether this respondent changed. These variables are derived using the wave specific and interview specific (i.e. main/second/history) respondent identifier.

Box 6 Syntax for Index file derived variables 7 to 9

```
7. compute mainint=-2.
if ((cw1mainres>0 and cw2mainres>0 and cw3mainres>0) and (cw1mainres=cw2mainres)
and (cw2mainres=cw3mainres)) mainint=0.
if ((cw1mainres>0 and cw2mainres>0 and cw3mainres>0) and
((cw1mainres~=cw2mainres)|(cw2mainres~=cw3mainres))) mainint=1.
if ((cw1mainres>0 and cw2mainres>0 and cw3mainres<0) and (cw1mainres=cw2mainres))
mainint=0.
if ((cw1mainres>0 and cw2mainres>0 and cw3mainres<0) and (cw1mainres~=cw2mainres))
mainint=1.
if ((cw1mainres>0 and cw2mainres<0 and cw3mainres<0)) mainint=0.
if ((cw1mainres>0 and cw2mainres<0 and cw3mainres>0) and (cw1mainres=cw3mainres))
mainint=0.
if ((cw1mainres>0 and cw2mainres<0 and cw3mainres>0) and (cw1mainres~=cw3mainres))
mainint=1.
if ((cw1mainres<0 and cw2mainres>0 and cw3mainres>0) and (cw2mainres=cw3mainres))
mainint=0.
if ((cw1mainres<0 and cw2mainres>0 and cw3mainres>0) and (cw2mainres~=cw3mainres))
mainint=1.
if ((cw1mainres<0 and cw2mainres>0 and cw3mainres<0)) mainint=0.
if ((cw1mainres<0 and cw2mainres<0 and cw3mainres>0)) mainint=0.
if ((cw1mainres<0 and cw2mainres<0 and cw3mainres<0)) mainint=-98.
if switch=1 mainint=-91.
if mainint=-5 and w1mainres=-994 and w2mainres=-994 and w3mainres=-994 mainint=0.
value labels mainint
-98 'No MP Interview'
-91 'Not Applicable - person switched position within household across waves'
0 'No change in MP person number'
1 'MP has changed between waves'.
exe.
```

```
8. compute secint=-2.
if ((cw1secores>0 and cw2secores>0 and cw3secores>0) and (cw1secores=cw2secores) and
(cw2secores=cw3secores)) secint=0.
if ((cw1secores>0 and cw2secores>0 and cw3secores>0) and
((cw1secores~=cw2secores)|(cw2secores~=cw3secores))) secint=1.
if ((cw1secores>0 and cw2secores>0 and cw3secores<0) and (cw1secores=cw2secores))
secint=0.
if ((cw1secores>0 and cw2secores>0 and cw3secores<0) and (cw1secores~=cw2secores))
secint=1.
if ((cw1secores>0 and cw2secores<0 and cw3secores<0)) secint=0.
if ((cw1secores>0 and cw2secores<0 and cw3secores>0) and (cw1secores=cw3secores))
secint=0.
if ((cw1secores>0 and cw2secores<0 and cw3secores>0) and (cw1secores~=cw3secores))
secint=1.
if ((cw1secores<=0 and cw2secores>0 and cw3secores>0) and (cw2secores=cw3secores))
secint=0.
if ((cw1secores<=0 and cw2secores>0 and cw3secores>0) and (cw2secores~=cw3secores))
secint=1.
if ((cw1secores<=0 and cw2secores>0 and cw3secores<0)) secint=0.
if ((cw1secores<=0 and cw2secores<0 and cw3secores>0)) secint=0.
if ((cw1secores<=0 and cw2secores<0 and cw3secores<0)) secint=-98.
if switch=1 secint=-91.
if secint=-5 and w1secores=-994 and w2secores=-994 and w3secores=-994 secint=0.
value labels secint
-98 'No SP Interview'
-91 'Not Applicable - person switched position within household across waves'
0 'No change in SP person number'
1 'SP has changed between waves'.
exe.

9. compute HISint=-2.
if ((Cw1histres>0 and Cw2histres>0) and (Cw1histres=Cw2histres)) HISint=0.
if ((Cw1histres>0 and Cw2histres>0) and (Cw1histres~=Cw2histres)) HISint=1.
if (Cw1histres>0 and Cw2histres<=0) HISint=0.
if (Cw1histres<=0 and Cw2histres>0) HISint=0.
if (Cw1histres<=0 and Cw2histres<=0) HISint=-98.
if switch=1 HISint=-91.
if hisint=-5 and w1histres=-994 hisint=-994.
value labels HISint
-994 "New member of household at Wave four (inc boost cases)".
-98 'No HISTORY Interview'
-91 'Not Applicable - person switched position within household across waves'
0 'No change in HISTORY person number'
1 'HISTORY person number has changed between waves'.
exe.
```

C. Variable indicating Household Grid completion

Variable 10 in Box 7 has been derived to identify a longitudinal level response across Waves 1 to 4. This variable is based on whether the Household Grid has been completed at each specific wave – therefore if a Household Grid is only available at Wave 1 the variable *resps* will indicate that the longitudinal level response is Wave 1 only.

Box 7 Syntax for Index file derived variable 10

```
10. compute resps=-2.
if w1hhgrid>=0 and w2hhgrid>=0 and w3hhgrid>=0 and w4hhgrid>=0 resps=1.
if w1hhgrid>=0 and w2hhgrid<0 and w3hhgrid<0 and w4hhgrid<0 resps=2.
if ((w1hhgrid=-992|w1hhgrid=-971) and (w3hhgrid=-992|w3hhgrid=-971) and w4hhgrid=-971
and w2hhgrid>=0) resps=3.
if ((w1hhgrid=-993|w1hhgrid=-972) and (w2hhgrid=-993|w2hhgrid=-972) and w4hhgrid=-972
and w3hhgrid>0) resps=4.
if w1hhgrid<0 and w2hhgrid<0 and w3hhgrid<0 and w4hhgrid>=0 resps=5.
if w1hhgrid>=0 and w2hhgrid>=0 and w3hhgrid<0 and w4hhgrid<0 resps=6.
if w1hhgrid<0 and w2hhgrid>=0 and w3hhgrid>=0 and w4hhgrid<0 resps=7.
if w1hhgrid>=0 and w2hhgrid<0 and w3hhgrid>=0 and w4hhgrid<0 resps=8.
if w1hhgrid>=0 and w2hhgrid<0 and w3hhgrid<0 and w4hhgrid>=0 resps=9.
if w1hhgrid<0 and w2hhgrid>=0 and w3hhgrid<0 and w4hhgrid>=0 resps=10.
if w1hhgrid<0 and w2hhgrid<0 and w3hhgrid>=0 and w4hhgrid>=0 resps=11.
if w1hhgrid>=0 and w2hhgrid>=0 and w3hhgrid>=0 and w4hhgrid<0 resps=12.
if w1hhgrid>=0 and w2hhgrid>=0 and w3hhgrid<0 and w4hhgrid>=0 resps=13.
if w1hhgrid>=0 and w2hhgrid<0 and w3hhgrid>=0 and w4hhgrid>=0 resps=14.
if w1hhgrid<0 and w2hhgrid>=0 and w3hhgrid>=0 and w4hhgrid>=0 resps=15.
variable labels resps "DV: Longitudinal household level response".
value labels resps
1 'Response in all waves'
2 'Response in W1 only'
3 'Response in W2 only'
4 'Response in W3 only'
5 'Response in W4 only'
6 'Response in W1 & W2 only'
7 'Response in W2 & W3 only'
8 'Response in W1 & W3 only'
9 'Response in W1 & W4 only'
10 'Response in W2 & W4 only'
11 'Response in W3 & W4 only'
12 'Response in W1, W2 & W3 only'
13 'Response in W1, W2 & W4 only'
14 'Response in W1, W3 & W4 only'
15 'Response in W2, W3 & W4 only'.
exe.
```

Any enquiries regarding this document/publication should be sent to us at team.longitudinal@education.gsi.gov.uk

Department for Education