Institute of Education

UCL

Next Steps

Sweep 8 – Age 25 Survey

User Guide (Second Edition)

Edited by Lisa Calderwood

(With contributions from Darina Peycheva, Morag Henderson, Tarek Mostafa, Sarab Rihal and the Next Steps Team)

January 2018

CENTRE FOR LONGITUDINAL STUDIES

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Contents

1. Introduction

The Next Steps age 25 survey took place between August 2015 and September 2016. It was designed and managed by the Centre for Longitudinal Studies (CLS) at the UCL Institute of Education (IoE), and fieldwork was carried out by NatCen Social Research. It was funded by the Economic and Social Research Council. This is the eighth sweep of the study following on from seven prior sweeps of data collected annually between 2004 and 2010. The study was previously run by the Department for Education and known as the Longitudinal Study of Young People in England (LSYPE).

The Next Steps age 25 survey involved a sequential mixed-mode design. Participants were first invited to participate online, non-responders were then contacted by telephone and face-to-face afterwards.

The age 25 survey sample design was to contact all cohort members who had ever taken part in any of the previous sweeps of the study (except those who had given a clear refusal or are ineligible). A total of 15,531 cohort members were issued for fieldwork and interviews were completed with 7,707 cohort members, representing a 51% response rate. The majority (62%) of fully productive interviews were achieved via web, 9% via telephone, and 29% of interviews were achieved face-to-face.

Across all modes completion of the survey took on average 47 minutes. The mean interview length was longest in the face-to-face mode (57 minutes) and shortest in web (44 minutes). The interview length was on average 51 minutes over the telephone.

A full account of the study development and fieldwork procedures can be found in the Next Steps Age 25 Survey Technical Report and Appendices, produced by NatCen Social Research, which accompanies this data deposit.

This user guide provides information about the data arising from the Next Steps age 25 survey and supports the deposit of the data at the UK Data Archive.

The deposit also includes:

Next Steps Age 25 Survey Questionnaire

Next Steps Age 25 Survey Technical Report and Appendices

Next Steps Age 25 Survey Derived Variables Guide

All cases in the dataset are identified by a unique case identifier which is consistent across all datasets from the First Longitudinal Study of Young People in England (LSYPE) (see Section 2. History and background) available from the UK Data Service, so the data can be linked with the data collected in previous sweeps. **Note that the case identifier has been changed on both sweep 8 and prior sweep data. It is now 'NSID' which replaces the old case identifier 'surveyid'.** Questionnaire documentation and user guide relating to the first seven sweeps of the study are available from the UKDS website,

This guide was revised in January 2018 (second edition) to reflect the UKDS deposits of the unfolding bracket datasets (December 2017) and the Derived Variables Guide (October 2017). The main revision relates to: additional information under Sections 8.6 'Unfolding brackets' and 8.7 'Known issues and data cleaning. Other revisions include editing of labels under Section 5.3 'Derived Variables' and minor formatting changes.

2. History and background

2.1 History of Next Steps

Next Steps, previously known as the Longitudinal Study of Young People in England (LSYPE), follows the lives of around 16,000 people born in 1989-90. The study began in 2004 and included young people in Year 9 who attended state and independent schools in England. Following the initial survey at age 13-14, the cohort members were interviewed every year until 2010, when they were age 19-20, to map their journeys from compulsory schooling to university, training and, ultimately, entry into the labour market. Therefore questions over the past seven sweeps (2004 to 2010) have mainly focused on the educational and early labour market experiences of young people, but also included diverse information on aspects of their lives including social participation and attitudes, risky-, crime- and anti-social behaviours, health and wellbeing, family formation, and aspirations for the future. The survey data has also been linked to the National Pupil Database (NPD) records, including cohort members' individual scores at Key Stage 2, 3 and 4.

The data for the first four sweeps (2004 to 2007) of the study was collected via face-to-face interviewing, and included interviews with cohort members' parents to understand the family environment. From sweep 5 (2008) onwards, interviewing involved the young person only and used a sequential mixed mode approach. Cohort members could complete the interview online, over the telephone or face-to-face.

The first seven sweeps of the study (2004 to 2010) were funded and managed by the Department for Education (DfE). The DfE's remit for running the Next Steps study was around compulsory education and there was no intention at sweep 7 to continue research with the cohort.

2.2 Continuation of the cohort

Next Steps, however, is the only major national longitudinal study focussing on young people's transitions to adulthood and their pathways through the teenage years, and thus a strategically important data resource for UK social science. Continuing the cohort therefore provides a unique opportunity to increase our understanding of transitions out of education and into early adult life.

The Next Steps cohort also represented a major opportunity to fill a 30 year gap in the series of birth cohort studies in Britain since 1946. Cohort studies were started for those born in 1970 and 2000, and the Next Steps cohort represents the missing 'Millennials' of 1989-1990.

In 2013, funded by the Economic and Social Research Council (ESRC), the management of Next Steps was transferred to the Centre for Longitudinal Studies (CLS) at the UCL Institute of Education. The intention is for CLS to run the Next Steps cohort alongside the existing national birth cohorts and thus provide a general purpose public-use data resource, for both the scientific community and policy makers, as the cohort moves through the transition to independent adult life and beyond.

The content of the Age 25 survey, conducted in 2015/2016, was therefore broadened slightly away from its original focus on education with the aim of it to become a more multi-disciplinary research resource, and a wide range of data linkage consents were collected during this survey sweep (see Section 5.1 Questionnaire overview and timing).

In the transfer of the study from DfE to CLS, DfE approached all 16,122 cohort members who had ever taken part in the study (except a small number who had previously withdrawn)

with an opportunity to decline having their contact information passed on to CLS. Following the DfE opt-out mailing, 15,629 cases were transferred to CLS.

From sweep 2 to sweep 7, only those who participated at the previous sweep were included in the issued sample for the current sweep. This meant that a significant proportion of the sample approached at sweep 8 had not participated in the study for a number of years. 53% of cohort members approached for sweep 8 had last participated at age 19, and 14% had not participated since they were age 13 and had only done so on that one occasion.

To maximise sample quality, at sweep 8 CLS attempted to trace and re-contact everyone ever participated in the study, except those who had given an adamant refusal to be part of it or have been identified as ineligible (see Section 3 Sample and response). Reviving the cohort required a significant effort in every aspect of the study's approach with substantial work involved in maintaining contact – re-contacting and re-engaging the cohort with the study (see Section 4 Design and Development).

3. Sample and response

3.1 Original sample design (including sample boosts)

The Next Steps population consisted of young people in Year 9 in England in state and independent schools and pupil referral units in February 2004. Sample members were born between 1st September 1989 and 31st August 1990.

The sample design considered schools the primary sampling unit, with deprived schools being over-sampled by 50%. Of 892 selected schools, 647 state and independent secondary schools as well as pupil referral units participated in the study. Within selected schools, pupils from minority ethnic groups (Indian; Pakistani; Bangladeshi; Black African; Black Caribbean; and Mixed) were over-sampled to provide sufficient base sizes for analysis. The school and pupil selection approach ensured that, within a deprivation band and ethnic group, pupils had an equal probability of selection. In addition to the young person, a 'main' and a second parent were identified for interview in each wave up to and including sweep 4; from sweep 5 onwards, only the young person has been interviewed.

The issued sample for Sweep 1 was approximately 21,000 young people. A total of 15,770 households were interviewed in that initial wave, representing 74 per cent of the target sample, with both young people and their parents in scope to be interviewed. At Sweep 4, 352 ethnic boost interviews were added (352 Black Caribbean and Black African pupils, selected from the original school sample), taking the total number of cohort members who had taken part in the study up to 16,122.

A detail description of the sampling at the first seven sweeps of the study is available in the LSYPE User Guide to the Datasets: Wave 1 to Wave 7, at the UKDS website

3.2 Sample at sweep 8 (including eligibility criteria)

Following the initial wave of fieldwork and up until sweep 7 in 2010, only those who participated in the previous wave were included in the current survey. This led to a reduction in the overall sample to 8,682 at sweep 7, or 53% of the total sample of 16,122.

To improve the representativeness of the sample of young adults and maximise the sample size, at Sweep 8 CLS attempted to trace and re-contact everyone who had ever taken part in in the study (except a small number who had previously withdrawn).

Prior to transferring the study to CLS, DfE approached all the 16,122 cohort members (except a small number who had previously withdrawn) to ask for updated contact information and provide an opportunity to opt out of the study. The opt-out mailing involved an update of cohort members' contact details using the National Pupil Database (NPD) - containing records for all state school pupils in England, including their home address; and the Individualised Learner Record (ILR) - containing records of students in vocational education and training post-16.

In total, 15,629 cases were then transferred to CLS from DfE. Following the receipt of the sample, CLS further sought to update the cohort members' information using the NHS Central Register - a database of GP registrations held by the Health and Social Care Information Centre (now NHS Digital), also providing information on individuals who have died or have moved out of the country; the electoral roll, phone records and postal directories.

The individual records were continually updated following contact with the study participants, through the Next Steps website, social media (Facebook and Twitter), e-mails, telephone

calls or the return of change of address cards (enclosed in the opt-out mailing, conducted by DfE). A 'participant pack' mailing further sought to reintroduce the study and encourage cohort members to contact CLS with updated contact information.

Cohort members were not issued for fieldwork where they were known to be:

- In prison
- Deceased
- Outside the UK
- Identified by CLS as in the armed forces or as out of the survey for another reason.

Those outside the UK were technically ineligible during fieldwork, but would have been able to complete an interview online where an email address was available. During telephone interviewing, numbers outside the UK were not called (and from the second batch of fieldwork onwards, only productive at sweep 7 cases were issued to telephone fieldwork). Cohort members outside England were not contacted during the face-to-face fieldwork, although they remained eligible to the study (see Eligibility criteria per mode in Section 5.2.3 Eligibility of cases for each mode in the Next Steps Wave 8 Technical report).

Following exclusions for known ineligibility and adamant refusals, a total of 15,531 cohort members were issued for fieldwork at sweep 8.

3.3 Response at sweep 8

A total of 15,531 cohort members were issued for fieldwork, with 423 of those found to be ineligible during fieldwork. Interviews were achieved with 7,707 cohort members, representing a 51% response rate (among eligible cases). Of the total, 7,481 were fully productive (97%).

Table 1 below shows the overall response for the cases issued to fieldwork. It shows that the main response for non-response was non-location (untraced).

Table 1: Sweep 8 overall response for issued sample

Outcome group	Frequency	Percent
Productive	7707	49.6
Refusal	2426	15.6
No contact	1482	9.5
Untraced	2996	19.3
Other unproductive	496	3.2
Ineligible	424	0.3
Total	**15531**	**100.0**

4,909 interviews were achieved via Web, representing a 32% response rate. This includes Web interviews achieved during telephone and face-to-face fieldwork (4%). A total of 5,297 cases were issued to telephone, and 719 interviews achieved. This represented a 14% response rate or 5% of all cases issued. Of the 10,357 issued to the face-to-face mode, 2,220 were interviewed (representing a response rate of 22%, or 14% of all issued cases). (See Section 6 for further details of the survey response in the Next Steps Age 25 Survey Technical report).

The number of interviews in each mode in the above paragraph sums to more than the total number of interviews. This is due to double-counting of partial interviews which are then fully productive in a later mode. Table 2 below shows the number of fully and partially productive interviews by the final mode.

Table 2: Full and partial interviews by Mode

		Outcome		Total
		Fully productive	Partially productive	
Mode of interview	Web	4615	182	4797
	TEL	660	30	690
	F2F	2206	14	2220
Total		7481	226	7707

Table 3 below shows that over 70% of productive cases were last interviewed in Sweep 7. The remainder of cases were last interviewed at earlier sweeps.

Table 3: Number of interviews by wave of last participation

	Frequency	Percent
Wave 1	535	6.9
Wave 2	274	3.6
Wave 3	258	3.3
Wave 4	229	3.0
Wave 5	318	4.1
Wave 6	507	6.6
Wave 7	5585	72.5
*Missing**	*1*	*.0*
Total	**7707**	**100.0**

**1 HH not in wave/missing information*

4. Design and development

4.1 Overview

A number of organisations were involved in the development and delivery of the Next Steps Age 25 survey. The Centre for Longitudinal Studies (CLS) were responsible for the development of the survey, funded by the Economic and Social Research Council (ESRC).

The content of the Age 25 Next Steps questionnaire was developed in collaboration with a wide range of academics, data users and other stakeholders. The National Centre for Social Research (NatCen) assisted CLS with the development of the instrumentation, conducted the fieldwork and carried out initial data preparation (including coding and post field editing where applicable) and documentation.

Ethical approval for the study was secured by CLS from the NHS Research Ethics Committee (NRES) – REC Reference 14/LO/0096.

4.2 Development work

Reviving the Next Steps cohort involved considerable effort in every aspect of the study's design, aimed at re-establishing contact and re-engaging participants, maximizing participation and optimizing the participant experience during the survey that they remain engaged with the study over time.

4.2.1 Communication with participants

Given the length of time since the previous interview, and the expected high proportion of movers in this age group, tracing but also finding the best way to re-engage cohort members with study was vital. Significant effort was therefore put to trace - in office and via administrative data records - and find most updated contact information (see Section 3.2 Sample at sweep 8; 2.2 Pre-fieldwork tracing in Next Steps Age 25 Survey Technical Report).

Re-engaging the cohort was vital and a comprehensive communication strategy thus underpinned the fieldwork approach. Considerable attention was given on the development of coordinated mailings and other communications to cohort members; development of a new participant-facing study website; a social media campaign; development of key messages and content in a range of media that could be used to engage the cohort and encourage participation (see Section 5.3.1 Communications with participants in Next Steps Age 25 Survey Technical Report). As part of this strategy, CLS sought to develop a strong brand - a new logo and visual identity for Next Steps. The 'Next Steps' name was retained from previous waves as this was how the study was known to cohort members.

4.2.2 Special elements

The engagement of the cohort was further sought through the study design and instruments aimed at optimizing the participant experience during completion. The mixed-mode approach offered a range of options to cohort members to participate in the way they find most convenient and appropriate (online, over the telephone or face-to-face). Attention was given to adapting questions to be mode-appropriate, while ensuring that mixing modes would not threaten the comparability of the data. A number of instrument specific features were implemented aimed at improving the data collection process such as a keyword look-up for coding occupation; an Event History Calendar (EHC); adaptations to sensitive questions for completion in the telephone mode; a progress bar for Web; embedded videos in Web and face-to-face (see Section 3.5 Special elements in Next Steps Age 25 Survey Technical

Report). These elements were tested prior to the main stage of fieldwork to establish the most effective design and implementation.

4.2.3 Asking for administrative data linkage consent

Data linkage consent was a major part of the study, and the communication of the consent request, as well as the broader process and protocol, were a key area for testing prior the main data collection stage. Exploratory qualitative work tested its general acceptability to participants and contextual issues; all participant materials and operational procedures were tested in the study pilot and approved by data holders and ethical committees (see section 6. 2 Consent process).

4.3 Pilot

Pilot fieldwork took place in October/November 2014 with a fresh sample of 212 individuals aged 23-27 recruited by the interviewers. The pilot used a concurrent mixed mode design, with 90 cases issued to Web, 79 to telephone and 43 for face-to-face fieldwork.

The purpose of the pilot was to test the questionnaire length, the functioning of the survey instrument and questions across modes, and to assess the protocols for seeking consents for data linkage. The pilot was also aimed at gaining feedback on the participant engagement strategies and fieldwork materials.

The highest response during the Pilot was achieved via the face-to-face mode (65%), with the face-to-face interview taking on average the least time to complete (57 minutes). Considerable cuts to the questionnaire followed the study Pilot (see Section 4.3 Piloting in Next Steps Age 25 Survey Technical Report for findings on special elements).

4.4 Soft launch and mainstage fieldwork

Next Steps Age 25 Survey data collection took place from August 20th 2015 until September 25th 2016 (See Section 5.8 Timing of fieldwork in Next Steps Age 25 Survey Technical Report). A sequential multimode design was used involving online (Web), telephone (Tel) and face-to-face (F2F) data collection. Incentives of £10 or £20 were offered to participants conditional on completing the survey. A £20 incentive was given for completion of the web survey within the first three weeks and £10 thereafter. This approach was tested experimentally in the soft launch (see Section 5.3.2 of the Next Steps Age 25 Survey Technical Report).

To make fieldwork more manageable, issued sample was divided into four batches, released to the field in sequence. Batch 1 was designated as a ‘soft launch’ sample to enable testing the survey processes and provide evidence of likely response at each mode.

All cohort members were initially invited to complete the survey online (and this mode of completion remained open throughout the fieldwork period). They had three weeks to complete the web survey before being issued to telephone. Over the three weeks and following an advance mailing, cohort members were sent three emails, two postal and two text reminders, if they had not started the web survey at the time of dispatch. Break-off reminders, via e-mail and text message, were sent to participants who have started or partially completed the survey (see Section 5.3.1 Communications with participants in Next Steps Age 25 Survey Technical Report).

Unproductive web cases, who had a valid telephone number and (from Batch 2 onwards) have participated in sweep 7, were contacted by phone three weeks after web started (see Section 5.2.3 Eligibility of cases for each mode in Next Steps Age 25 Survey Technical

Report). All (40) telephone interviewers attended a one day project briefing before starting work on the study.

Once all telephone numbers had been attempted for the cohort member without making contact, telephone reminder texts were sent to inform cohort members that a telephone interviewer was attempting to contact them. An appointment reminder text message was sent before agreed interview appointments. From Batch 2, an email was sent after telephone stage was closed, but before face-to-face, to again encourage the option of completing the survey online.

A maximum number of 10 calls was set over the period to avoid negative reaction at face-to-face. Calls were made weekday day time, week day evenings and weekends. The mean number of telephone calls per productive interview was 5.4. Tracing using stable contact telephone number (where held) was conducted by the interviewer once all numbers held for the cohort member had been exhausted.

Over 10,000 cases were issued for face-to-face fieldwork as a third and final mode of data collection. An additional advance letter was sent out to all cases to inform them that an interviewer would call at their address in the next few weeks. Prior to commencing work on the study, all (214) face-to-face interviewers attended a one day project briefing.

First contact during face-to-face fieldwork was face-to-face. Interviewers were required to make a minimum of six face-to-face calls including a minimum of one call each at evenings and weekends and three evening/weekend calls in total, with no maximum number of calls set. The average number of face-to-face calls per productive interview was 3.8.

Interviewers were provided with a study-branded message card to leave a message. Text and/or emails messages were sent by interviewers to confirm appointments.

The Next Steps Age 25 Survey fieldwork accommodated extended tracing attempts and multiple reallocations of work (see Sections 5.4 Contact and tracing and 5.5 Reallocation and reissuing in Next Steps Age 25 Survey Technical Report). A total of 7,052 (45%) cohort members were identified as having moved from the original address at the start of the live fieldwork. Of those new addresses were collected for 4,119 (58%) and interviews achieved with 947 cases amongst movers (see Section 6.3 Tracing response by mode in Next Steps Age 25 Survey Technical Report). 10% of cases issued for face-to-face fieldwork were reissued to interviewers to be worked again with an overall response rate of 7%.

5. Questionnaire content

5.1 Questionnaire overview and timing

A wide range of questions have been asked over the past seven sweeps. They have mainly focused on the educational experiences of young people, but information has also been collected about their employment, economic circumstances, family life, physical and emotional health and wellbeing, social participation and attitudes.

The content of the survey at age 25 was broadened with the aim of it becoming a more multi-disciplinary research resource, and covered the following topics:

Table 4: Questionnaire content at Wave 8

Number	Module	Content	Summary of sweep 8 topics covered in sweeps 1 to 7
1	Household relationships	• Current relationship • Previous cohabiting relationships • Children • Childcare • Non-resident children • Non-resident parents • Other household members	Household (relationships) data have been collected in all previous study waves. This information was provided by the cohort member from Sweep 5 onwards. (In Sweep 8, data is collected on past cohabiting relationships, dating back to September 2006).
2	Housing	• Current housing • Previous housing (summary)	Housing history data, if the cohort member has moved out of their parent's home, was not obtained in previous study waves. (In Sweep 8, summary data is therefore collected about the different addresses the study members have lived in since they were 16, if other than the parents' home).
3	Employment	• Current Activity • Current Employment • Second job	Data on current economic activities and activity history,

		• Prospective employment (for unemployed) • Activity history • Employment Details for first job after September 2006 (aged 16) • Employment support • Work attitudes • Partner employment	has been collected in Sweep 4 to 7 (previous activity data in Sweep 8 was therefore obtained back to the time of the last interview and no earlier than September 2006). Job search related information was collected in Sweep 5 to 7 and attitudes to work data in Sweep 4 to 6.
4	Finance	• Current pay/salary main job • Pay from second job • Income from other jobs • Partner income • Benefits • Income from other sources • Household income • Pensions • Debt	Income and benefits data was collected from sweep 4 to 7.
5	Education and Job Training	• Job training • Education since previous interview/September 2006 • Current education • Fees • Partner education	Jobs training, as well as academic and vocational qualifications data, was obtained in Waves 4 to 7 (this information in Wave 8 was therefore obtained back to the time of the last interview and no earlier than September 2006). Higher education data was collected in Sweep 5 to 7, and questions on attitudes to higher education asked in Sweeps 6 and 7. Education maintenance allowance (EMA) information was collected in Sweeps 3 to 5.
6	Health and Wellbeing	• General health • Height and weights	Health and disability data was

		• Exercise • Sleep • Diet • Accidents and Injury	collected in Sweeps 2, 3, 4, 6 and 7; and mental and emotional health data in Sweeps 2, 4 and 7. Sport frequency information was obtained in all previous study waves apart from Sweep 5.
7	Identity and Participation	• Ethnic Group • Religion • Social Networks • Trust • Risk • Patience • Meritocratic beliefs • Adult identity • Leisure • Politics • Social Media	Ethnicity of the cohort member was recorded at the baseline study wave and religion collected from Waves 1 to 5. Social attitudes data has been collected in Sweep 5. Participation in all study waves apart from Sweep 6 and Use of leisure time in Sweep 1 to 4.
8	Self-completion	• Gender identity • Locus of control • Overall life satisfaction • GHQ-12 • Self-harm • Crime and harassment • Drinking behaviour • Smoking behaviour • Drugs • Bullying • Sexual behaviour • Pregnancy history	Antisocial activities data, bullying, contact with police have been obtained in all previous study waves. Overall life satisfaction was measured in Sweep 7, Locus of control in Sweep 2, 4 and 7. The GHQ-12 questionnaire embedded in Sweep 2 and 4. Sexuality and sexual experience, and pregnancies data was collected in Sweeps 6 and 7.
9	Data Linkage	Asking for consent to link to records held by the: • National Health Service (NHS) • Department of Work and Pensions (DWP)	Participating in the Next Steps study at Wave 1 was contingent on consent to link to the National Pupil

		• Her Majesty's Revenue and Customs (HMRC) • NI number • Department for Business Innovation and Skills (BIS) • Higher Education Statistics Agency (HESA) • Universities and Colleges Admission Service (UCAS) • Department for Education (DfE) • Student Loans Company (SLC) • Ministry of Justice (MOJ)	Database records. Since then data linkage consents have been obtained from cohort members in Sweep 4 to 7. These administrative data requests have been limited to DWP records linkage. In addition BIS records linkage request was asked in Sweep 7. However, the restrictive nature of the consent wording has meant that these consents were not enacted.

Note: A summary of the questionnaire content at the first seven waves of the study is available in the LSYPE User Guide to the Datasets: Wave 1 to Wave 7, at the UKDS website,

Interviewing, across all modes, took on average 47 minutes, with the longest interviewing being the face-to-face - 57 minutes - and Web being the shortest - 42 minutes. The modules that took the longest to complete were the Household relationships and Employment – on average 8 and 10 minutes, followed by the module with sensitive questions - 7 minutes on average.

Considerable effort was expended in the development of the data collection tools to minimise the risk that mixing modes would threaten the comparability of the data. Special attention was therefore given to adapting the questions to be mode-appropriate, while minimising the variance in question wording across modes. For this reason, variations between questions and across modes were modest and mostly limited to variations in the interviewer instructions to show a card or read out, and variations in question wording to assist online self-completion. For example, in the telephone mode, where all participants were asked the sensitive questions on drug use, sexual identity and pregnancy history over the phone, to avoid others in the household overhearing responses, interviewers read options out in full and have participants say 'yes' or 'no' at each option rather than ask them to say the answer out loud (see Interviewers' instructions in Module 8: Self-Completion in the Next Steps Age 25 Survey Questionnaire).

5.2 Special features

A number of special features were embedded in the data collection instruments aimed at improving the quality of the data collected.

One such special feature was the Event History Calendar (EHC). This tool created a visual timeline of the participant's life as life events (marriages, cohabitations, changes in employment status, address changes) were entered into the questionnaire. As the participant answers a particular question (for example when they started living at a particular

address), the calendar automatically updates to display the event in relation to their age, the date, and other events they had already coded.

A text-based search and coding system using the detailed four-digit standard occupation code frame (SOC2010) as a look-up file was added to the questionnaire, to reduce the need for office-based coding. This enabled participants or interviewers to enter key words to search for the occupation code that was most appropriate to them (for example if 'secondary school teacher' was typed, a short list of related codes was obtained from which to select).

A relatively small amount of feed forward data (participant information from previous survey) was used at various points in the questionnaire for routing and checks. These included confirmation of date of birth, contact details, and historical information about relationships and economic activity.

5.3 Derived variables

A number of derived variables have been produced based on the questionnaire data and are listed below. The majority of these can be found in a separate derived variables dataset and detailed documentation on their derivation can be found in the Derived Variables Guide.

Household Relationships		
	Variable Name	**Variable Label**
	W8DHSIZE	DV: Number of people currently living in household (inc CM)
	W8DCHNO	DV: Number of children currently living in household
	W8DCHOWNNO	DV: Number of own children currently living in household
	W8DCHPARNO	DV: Number of children of CM's current or previous partner in household
	W8DFATHER	DV: Whether CM's father in household
	W8DMOTHER	DV: Whether CM's mother in household
	W8DMARSTAT	DV: Legal marital status
	W8DNRAGE	DV: Age of CM's partner
	W8DPARAGE	DV: Age of biological parent
	W8DNCHRAGE	DV: Age of own children
	W8DAGEYCH	DV: Age of CM's youngest child
	W8DCHNO4	DV: Number of own children between 0 and 4
	W8DCHNO11	DV: Number of own children between 5 and 11
Housing		
	Variable Name	**Variable Label**
	W8DTIMAD	DV: Time at current address (months)
Employment		
	Variable Name	**Variable Label**
	W8DACTIVITYC	DV: Current activity of CM - backcoded
	W8DWRK	DV: Whether CM currently employed - backcoded
	W8DEMPSZ	DV: Employment status/size of organisation for cohort member
	W8DNSSEC17	DV: NS-SEC (2010) (Full operational categories)
	W8DNSSEC13	DV: NS-SEC (2010) 13 (Combined operational categories)

	W8DNSSEC8	DV: NS-SEC (2010) 7 (Analytic classes)
	W8DNSSEC5	DV: NS-SEC (2010) 5 (Analytic classes)
	W8DWRKP	DV: Whether CM's partner currently employed - backcoded
	W8DDACTIVITYP	DV: Current activity of CM's partner
	W8DWRKCP	DV: Combined labour market status - backcoded
Finance		
	Variable Name	**Variable Label**
	W8DBENE	DV: Whether cohort member or partner receives any benefits
	W8DBENE2	DV: Whether cohort member or partner receives any benefits (incl extra split)
Education		
	Variable Name	**Variable Label**
	W8DANVQH	DV: Highest NVQ level from an academic qualification in 2015
	W8DHANVQH	DV: Highest NVQ level from an academic qualification to 2015
	W8DDEGP	DV: Whether achieved first degree or higher
	W8DRUSSELL	DV: Whether degree awarded by Russell Group University
	W8HESUBGROUP	DV: Degree subject (grouped)
Health		
	Variable Name	**Variable Label**
	W8DBMI	DV: Body mass index
	W8DBMICA	DV: Body mass index category
Identity		
	Variable Name	**Variable Label**
	W8DETHN6	DV: Ethnic group - 6 category census class
	W8DETHN8	DV: Ethnic group - 8 category census class
	W8DETHN11	DV: Ethnic group - 11 category census class
	W8DETHN15	DV: Ethnic group - Detailed
	W8DETHNP6	DV: Ethnic group of CM's partner - 6 category census class
	W8DETHNP8	DV: Ethnic group of CM's partner - 8 category census class
	W8DETHNP11	DV: Ethnic group of CM's partner - 11 category census class
	W8DETHNP15	DV: Ethnic group of CM's partner - Detailed
	W8DRELIG7	DV: Religion - 7 category
Self-completion		
	Variable Name	**Variable Label**
	W8DCANEVER	DV: Whether CM has ever tried cannabis
	W8DSEXEVER	DV: Whether CM has ever had sex
	W8DPREG	DV: Whether CM has ever been pregnant

5.4 Income

5.4.1 Questionnaire measures

Income was collected in sweep 8 of Next Steps using five separate banded questions. It was administered in this way, rather than with a single, longer list of income bands, to facilitate administration over the phone. The first question gives respondents a choice between four bands, and the four remaining questions subdivide each band into four finer bands. In total the scale consists of 16 bands.

The questions explicitly cover different income sources like earnings, state benefits, and other sources of income. The exact wording of the income questions is provided in the Sweep 8 questionnaire.

Income was missing for 9.4% of the 7,707 respondents in wave 8. The reasons for income item non-response are: refusal, don't know, not applicable.

5.4.2 Imputation of missing and continuous income from banded data

Income was imputed using interval regression (Stewart 1983). This method allowed us to impute a continuous value within a band, rather than assuming that all cases in a band had the same midpoint income. This was achieved using Stata's INTREG command (StataCorp 2007; Conroy 2005). INTREG fits a model of y = [dependent variable 1, dependent variable 2] on independent variables where the dependent variable 1 was the log lower income band and dependent variable 2 was log upper income band. The INTREG procedure also allowed us to impute all missing values on the income questions.

Note that the left-hand-side bound for the lowest band is £0 per week and the right-hand-side bound for the top band was fixed at £1,700 per week. The predictors are shown below.

5.4.3 Predictors of income in Sweep 8

Family circumstances and cohort members' (CM) characteristics in sweep 1:

1. Highest qualification held by main parent (sweep 1)
2. Employment status of main parent (sweep 1)
3. Social status NS-SEC of the family (sweep 1)
4. Marital status of main parent (sweep 1)
5. Cohort member's gender (sweep 1)
6. Cohort member's ethnic group (sweep 1)
7. Whether cohort member ever identified as having special educational needs sweep 1)
8. Government office region (GOR sweep 1)

Cohort member's circumstances in sweep 7:

1. Housing tenure (sweep 7)
2. Current activity including education and employment (sweep 7)
3. Whether cohort member ever tried cannabis (sweep 7)
4. Month of interview (sweep 7)
5. Interview mode (sweep 7)

Missing data for the predictor variables due to non-monotone non-response or item missingness were imputed as described in Appendix A.

5.4.4 Income derived variables

Variable name	Variable label
W8DINCW	DV: Continuous weekly income
W8DINCB	DV: Banded weekly income

5.5 Scales

The Next Steps Age 25 Survey included several established scales which are listed below. Overall scores for each scale have been derived and included within the data deposit. Further details regarding the derivation of the scores can be found in Derived Variables Guide.

Health module: Long-lasting Health Conditions and Illnesses: Impairments and Disability (ONS, 2015)

The Age 25 Survey included a sub-set of the ONS harmonised set of questions on Long-lasting Health Conditions and Illnesses including Impairments and Disability. The three items listed below are used to derive variables indicating whether cohort members are disabled using the Equality Act 2010 definition (W8DDISEA) and whether they have a long-standing illness or condition using the European Union's Statistics on Income and Living Conditions (EU-SILC) definition (W8DDISEU) (ONS, 2015). W8DDISEA identifies individuals as disabled or not, W8DDISEU identifies individuals as having no long-standing health condition, having a condition which hampers daily activities to an extent and having a condition which severely hampers daily activities.

Variable name	Variable label
W8LOIL	Has longstanding illness
W8LOLM	Reduced day-to-day activities as result of longstanding illness
W8LOLP	Length of time day-to-day activities affected by longstanding illness
W8DDISEA	DV: Disability classification Equality act (2010)
W8DDISEU	DV: Disability classification EU-SILC

According to the Equality Act 2010 definition, a cohort member is considered to be disabled if they report a longstanding illness (W8LOIL), and have a reduced ability to carry out day-to-day activities as a result of their illness (W8LOLM).

According to the EU-SILC definition, a cohort member is considered to be disabled if they report a longstanding illness (W8LOIL), have a reduced ability to carry out day-to-day activities as a result of their illness (W8LOLM), and this reduced ability has lasted for more than 6 months (W8LOLP). This variable also distinguishes between those that are disabled to some extent, and those that are severely hampered (from W8LOLM).

Modified versions of the above items have been asked to cohort members in Waves 4, 6 and 7. Parents' reports have been collected in Waves 1 and 2.

Identity module: Adult Identity Resolution Scale (Côté, 1997)

The Adult Identity Resolution Scale (AIRS) is subscale of the Identity Stage Resolution Index (ISRI).

Variable name	Variable label

W8ADULT0A	Whether reached adulthood: You consider yourself to be an adult
W8ADULT0B	Whether reached adulthood: You feel respected by others as an adult
W8ADULT0C	Whether reached adulthood: You feel you have matured fully

Self-completion module: Locus of control, GHQ-12, and Drinking behaviour

Locus of control (Lefcourt, 1991)

The four items below are used to derive a variable to indicate the extent to which participants believe that they have control over events in their lives.

Variable name	Variable label
W8LOCUS0A	Locus of control: If someone is not a success in life, it is usually their own fault
W8LOCUS0B	Locus of control: I can pretty much decide what will happen in my life
W8LOCUS0C	Locus of control: How well you get on in this world is mostly a matter of luck
W8LOCUS0D	Locus of control: If you work hard at something you'll usually succeed
W8DLOCUS	DV: Locus of control scale (Lefcourt,1991): overall score

The cohort members' total score on the locus of control scale was derived by summing the responses to the locus of control questions (W8LOCUS0A, W8LOCUS0B, W8LOCUS0C and W8LOCUS0D) to generate a total score ranging from 4 to 16. A low value of 4 to 7 indicates an internal locus of control, a score ranging between 8 and 11 indicates either a moderate internal or moderate external locus of control, and a score between 12 and 16 suggests external locus of control.

These items have also been asked to participants in Waves 7, 4 and 2.

GHQ-12 General Health Questionnaire (GHQ-12) (Goldberg & Williams, 1988)

The General Health Questionnaire (GHQ) was used as a screening tool of probable mental ill health. The 12 item screening instrument measures general, non-psychotic and minor-psychiatric disorders; and concentrates on the broader components of psychological ill health and characteristics as general levels of happiness, depression and self-confidence.

Each of the 12 GHQ items, six positively and six negatively phrased, are rated on a four-point scale to indicate whether symptoms of mental ill health are 'not at all present', present 'no more than usual', present 'rather more than usual' or present 'much more than usual'. Using the standard GHQ coding method (0-0-1-1), we assigned a score of zero for the first two responses above, and a score of 1 for the third and fourth responses to obtain a total GHQ-12 score. The maximum score for any individual study participant is therefore 12.

Variable name	Variable label
W8GHQ12_1	GHQ12: Concentrate on what doing
W8GHQ12_2	GHQ12: Lost sleep over worry
W8GHQ12_3	GHQ12: Playing a useful part in things
W8GHQ12_4	GHQ12: Capable of making decisions
W8GHQ12_5	GHQ12: Constantly under strain
W8GHQ12_6	GHQ12: Can't overcome difficulties

W8GHQ12_7	GHQ12: Enjoy day to day activities
W8GHQ12_8	GHQ12: Face up to problems
W8GHQ12_9	GHQ12: Unhappy or depressed
W8GHQ12_10	GHQ12: Losing confidence in self
W8GHQ12_11	GHQ12: Thinking of self as worthless
W8GHQ12_12	GHQ12: Reasonably happy
W8DGHQSC	DV: General Health Questionnaire (GHQ12) score (Goldberg & Williams,1988)

The cohort member's score on the General Health Questionnaire 12 point scale (GHQ12) was derived by summing responses to the twelve GHQ12 questions (W8GHQ12_1 to W8GHQ12_12). As in previous waves, this was scored according to the 0-0-1-1 method, in which the first two possible responses to each question were assigned a value of 0 and the third and fourth responses with a value of 1, resulting in a maximum possible score of 12 for this variable. A higher score on this scale indicates a greater likelihood of mental ill health.

The 12 GHQ items have also been asked at Waves 4 and 2.

Alcohol Use Disorders Identification Test Consumption (AUDIT C)

The AUDIT-C was used to capture alcohol consumption, problems and dependency. Responses to the 3 questions below are scored from 0 to 4 giving a maximum score of 12 (NS8AUDIT). Scores of 5 or more are considered AUDIT-C positive and associated with increasing or higher risk drinking.

Variable name	**Variable label**
W8AUDIT1	How often has a drink containing alcohol
W8AUDIT2	How many drinks containing alcohol has on a typical day of drinking
W8AUDIT6	How often had six or more drinks on one occasion in the past year
W8DAUDIT	DV: Alcohol Use Disorders Identification Test Consumption (AUDIT-C) scale

The cohort member's score on the Alcohol Use Disorders Identification Test Consumption (AUDIT-C) scale was derived by summing the responses to the AUDIT-C questions (W8AUDIT1, W8AUDIT2, and W8AUDIT6).

6 Data Linkage

6.1 Asking for administrative data linkage consent

In Module 9 of the questionnaire, Next Steps participants were asked for consent to link their survey answers to nine different administrative data sources, held by a number of different government departments and non-governmental bodies:

- Health records, held by the NHS, including Primary Care data - covering visits to family doctor and other health professionals, and Hospital Episode Statistics (HES) - covering admissions and attendance at hospital;
- Records about school participation and attainment, and pupil characteristics, kept by the Department for Education;
- Records about participation in further education and attainment, kept by the Department for Business Innovation and Skills;
- Records covering university participation and attainment, held by the Higher Education Statistics Agency (HESA);
- Records covering higher education applications and offers, held by the Universities and Colleges Admissions Service (UCAS);
- Records covering payments of student support, held by Student Loans Company (SLC);
- Information on benefit and employment programs, kept by Department for Work and Pensions (DWP);
- Information on employment, earnings, tax credits, occupational pensions and National Insurance Contributions, kept by Her Majesty's Customs and Revenue (HMRC);
- Respondents who consented to either DWP or HMRC linkage were also asked for their National Insurance number (NINO).
- Police National Computer (PNC) records covering arrests, cautions and sentences, held by the Ministry of Justice;

All participant materials and operational procedures involved in collecting data linkage consent were tested in exploratory qualitative work and the study pilot, and approved by data holders and ethical committees prior the main stage data collection.

6.2 Consent process

Data linkage was a very important part of the study and considerable effort was expended in developing an approach that would maximise consent (see 4.2 Data linkage consent process in Next Steps Age 25 Survey Technical Report). Asking for consent was particularly challenging in web surveys, in which there is no immediate support from an interviewer for clarification of questions and reassurance of one's privacy at the time of the interview. The web, however, offers other opportunities such as better content visualisation, likely to help participants absorb key messages quickly; or quick access to other useful sources of information – which were embedded in the Web interview. These included a video to inform participants about the data linkage process, and help screens added to each of the pages of consent questions to allow participants to access more detailed information (e.g. 'Which records would you like to add?', 'What do these records include?', 'Why is it helpful to add this information?').

Following the pilot and qualitative testing, a three stage process (pre, during and post interview) was followed to elicit informed consent.

A data linkage leaflet was included in the advance mailing, sent at the start of each batch of fieldwork. It gave information on the purpose, types, value and process of data linkage, and encouraged study members to contact the study team with any questions they might have.

During the interview, and following an introduction page, consents were recorded directly into the survey instrument. In the online survey, participants recorded their consent at questions within the self-completion instrument. In the telephone and face-to-face modes, consent was provided verbally by participants and recorded by the interviewer.

All participants were then sent a confirmation of their consents (as part of their 'Thank you' letter) by post or email.

6.3 Achieved consent rates

The level of consent was higher for the telephone (90%) and face-to-face (89%) modes compared to the online (69%). Overall, the level of any consent to any linkage was 77%, with higher levels of consent for education (70%) and lower for economic records – HMRC (57%) and DWP (59%).

Table 5: Consent to data linkage by mode of data collection

	Web		Tel		F2F		Total	
	N	%	N	%	N	%	N	%
Total asked consent questions (including partials reaching this point in the questionnaire	**4635**	**100%**	**660**	**100%**	**2207**	**100%**	**7502**	**100%**
Any consent	3190	69%	592	90%	1967	89%	5749	77%
NHS	2558	55%	505	77%	1839	83%	4902	65%
Education	2806	61%	564	85%	1893	86%	5263	70%
UCAS	2699	58%	555	84%	1832	83%	5086	68%
Student Loans Company	2327	50%	491	74%	1688	76%	4506	60%
HMRC	2147	46%	469	71%	1676	76%	4292	57%
DWP	2230	48%	505	77%	1727	78%	4462	59%
MoJ	2526	54%	513	78%	1740	79%	4779	64%
NINO	2039	44%	307	47%	1420	64%	3766	50%
No consent	1445	31%	68	10%	240	11%	1753	23%
Consent withdrawn	0	0%	0	0%	2	0%	2	0%
Other no consent	1445	31%	68	10%	238	11%	1751	23%

**Source Next Steps Age 25 Survey Technical report*

6.4 Linked data deposit and documentation

Separate documentation will support the deposit of any subsequently deposited linked data. Further documentation (in the form of working papers) will be produced to provide more detail on data linkage.

7 Response and weights

7.1 Response patterns

As with any longitudinal survey, Next Steps is subject to attrition. Attrition takes place when respondents drop out of the survey over time. This leads to two problems: (i) a reduction in sample size and loss of statistical power, and (ii) bias in sample composition. Sample bias arises when the likelihood of dropping out from the survey is correlated with the socio-demographic characteristics of the respondents. In this case, the survey will lose a particular type of respondents (e.g. disadvantaged families, ethnic minorities, etc) and the sample will no longer be representative of the population it was drawn from. However, there are statistical methods to deal with this, so as to ensure the remaining sample recovers (under reasonable assumptions) population parameters.

This section examines attrition up to sweep 8 (age 25) of Next Steps and presents the procedures used in the construction of the sweep 8 attrition weights. For more information on the construction of weights in the previous sweeps of Next Steps visit the following page (see webpage).

In Table 6, the proportion of productive and unproductive cases are presented. The table shows that the proportion of productive cases decreased over time from 97.8 per cent in wave 1 to 47.8 per cent in sweep 8.

Table 6: Productive and unproductive cases in all Next Steps sweeps

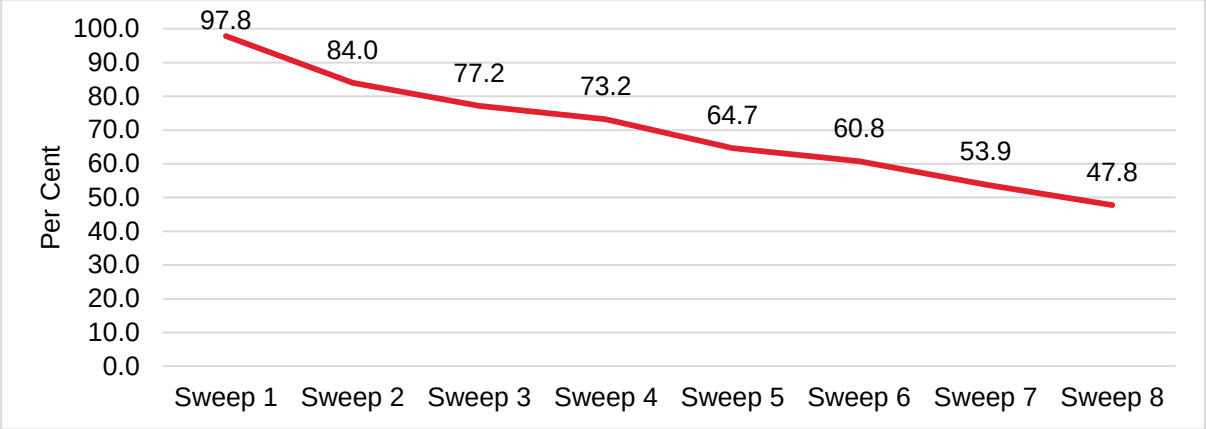
Sweeps	**NS1**		**NS2**		**NS3**		**NS4**	
Age in years	**13/14**		**14/15**		**15/16**		**16/17**	
	Freq.	**%**	**Freq.**	**%**	**Freq.**	**%**	**Freq.**	**%**
Productive	15,770	97.8	13,539	84.0	12,439	77.2	11,801	73.2
Unproductive	352	2.2	2,583	16.0	3,683	22.8	4,321	26.8
Total	16,122	100.0	16,122	100.0	16,122	100.0	16,122	100.0

Sweeps	**NS5**		**NS6**		**NS7**		**NS8**	
Age in years	**17/18**		**18/19**		**19/20**		**25**	
	Freq.	**%**	**Freq.**	**%**	**Freq.**	**%**	**Freq.**	**%**
Productive	10,430	64.7	9,799	60.8	8,682	53.9	7,707	47.8
Unproductive	5,692	35.3	6,323	39.2	7,440	46.2	8,415	52.2
Total	16,122	100.0	16,122	100.0	16,122	100.0	16,122	100.0

Note: The total number of Next Steps respondents ever interviewed is 16,122.

Figure 1 presents the proportion of productive cases in Next Steps in all sweeps. The figure shows that the sample decreased by 50% by the time of the sweep 8 survey.

Figure 1: Proportion of cases productive in all Next Steps sweeps



In Table 7 we look at different response patterns. Table 7 shows that 33.7 per cent of all respondents participated in all eight waves of Next Steps. In contrast, 16.7 per cent have interrupted response patterns (i.e. non-monotone response). In other words, they participated in a number of waves, and then dropped out before participating again in subsequent waves. The majority of these non-monotone cases are those who had dropped out at earlier waves and came back into the achieved sample at sweep 8. Non-monotone cases include the respondents in the top-up sample who came into the study at sweep 4 in addition to those with interrupted response patterns. The extent of non-monotone response is limited because up to sweep 7 Next Steps relied on a policy of issuing only cases which were productive on the previous wave with very few exceptions. Finally, 49.6 per cent of all respondents have monotone response patterns. That is, they participated in a number of waves before dropping out for all subsequent waves.

Table 7: Monotone vs. non-monotone response in Next Steps

Patterns	Freq.	Percent
Monotone	8,002	49.6
Non-monotone	2,694	16.7
All waves	5,426	33.7
Total	16,122	100.0

Table 8 shows the percentages of respondents participating in n sweeps (n=1…8). We see that 54.2 per cent of respondents participated in at least seven out of eight sweeps indicating that about half of the sample have almost complete records.

Table 8: Number of times productive up to Next Steps wave 8

Times productive	Freq.	Percent
One	1,752	10.9
Two	1,398	8.7
Three	1,063	6.6
Four	1,069	6.6
Five	1,048	6.5
Six	1,056	6.6
Seven	3,310	20.5
Eight	5,426	33.7
Total	16,122	100.0

7.2 Predicting response in sweep 8 and weights

The procedure used for predicting response in wave 8 is similar to the one used in the previous sweeps (see webpage). We estimated a logit model in which the dependent variable is binary (where 1 is for response and 0 is for otherwise) and the predictors are:

Family circumstances and cohort members' (CM) characteristics in sweep 1:

1. Highest qualification held by main parent (sweep 1)
2. Employment status of main parent (sweep 1)
3. Social status NS-SEC of the family (sweep 1)
4. Marital status of main parent (sweep 1)
5. Cohort member's gender (sweep 1)
6. Cohort member's ethnic group (sweep 1)
7. Whether cohort member ever identified as having special educational needs (sweep 1)
8. Government office region (GOR sweep 1)

Cohort member's circumstances in sweep 7:

1. Housing tenure (sweep 7)
2. Current activity including education and employment (sweep 7)
3. Whether cohort member ever tried cannabis (sweep 7)
4. Month of interview (sweep 7)
5. Interview mode (sweep 7)

Missing data for the predictor variables due to non-monotone non-response or item missingness were imputed as described in Appendix A.

Table A1 in the Appendix shows the odds ratios of the response logit model estimated using the 50 imputed datasets. The linear predicted values were generated from this model; then the predicted values were converted into predicted probabilities using an inverse logit transformation. The non-response weights for sweep 8 were constructed as the inverse of the predicted probabilities (Wooldridge 2007; White et al., 2011). The final weight for sweep 8 was constructed by multiplying the sweep 8 non-response weight with the final weight from sweep 7. The final weight from sweep 7 takes account of the design weight and of attrition weights in sweep 7 and in all previous sweeps. The sweep 8 weights were scaled to make their total equal to the productive sample size.

Weights variables:

Variable name	Variable label
W8FINWT	Weight: W8 Final Weight

When the Next Steps data are being used, researchers should use the weight corresponding to the most recent sweep included in their analysis. For instance, if they are using data from sweeps 1, 2, 5 and 8, the sweep 8 weights should be used.

In Table A2, the means, minima and maxima of the weights are presented. We note that the effectiveness of the response weights to correct for bias depends on the inclusion of the most important predictors of unit non response in the logit response model (Seaman and

White, 2011), as well as the plausibility of the MAR assumption in the interim Multiple Imputation of missing values in the variables that were used as predictors of response.

8 Datasets and data conventions

8.1 Datasets and format

The majority of the questionnaire data (modules 1 to 7) are deposited in a flat file format such that one record exists for each cohort member (CM). Responses to the self-completion (module 8) questionnaire are included in a separate dataset.

In addition, a number of hierarchical datasets are provided in which there are multiple records for cohort members. These datasets consist of responses to questions where the respondent is asked a set of questions which are repeated until no more information is required. Where the series of questions were not relevant to a particular cohort member they will have no records in that hierarchical dataset.

A feature of income or payment questions in this wave is the use of unfolding brackets where a respondent refuses or is unable to provide an exact answer. As these questions relate to a minority of respondents they have been placed in separate datasets.

Note that the case identifier used on the files is 'NSID' which replaces the old case identifier 'surveyid'.

The datasets are as follows:

Name	Contents	Structure	Identifier(s)
NS8_2015_Main_Interview	Modules 1 to 7	Flat	NSID
NS8_2015_Self_Completion	Module 8	Flat	NSID
NS8_2015_Partnerships	Relationship histories	Hierarchical	NSID, W8RELID
NS8_2015_Children	Details of children of CM	Hierarchical	NSID, W8CHID
NS8_2015_Household_Members	Details of members living in same household as CM	Hierarchical	NSID, W8HHMID
NS8_2015_Activity_History	Activities and Employment histories	Hierarchical	NSID, W8HISTID
NS8_2015_Benefits	Details of individual benefits received	Hierarchical	NSID, W8BENID
NS8_2015_Income_Unfolding_ brackets	Unfolding brackets questions for payments and income	Flat	NSID
NS8_2015_Benefits_Unfolding_ brackets	Unfolding brackets questions for benefits	Hierarchical	NSID, W8BENID
NS8_2015_Derived_variables	Derived variables	Flat	NSID
NS8_2015_Outcome	A summary of outcome codes for all issued cases	Flat	NSID

The Derived variables file contains geographical variables indicating the country and region of interview as follows:

Geography		
	Variable Name	**Variable label**
	W8DCTRY	DV: Country at interview
	W8DGOR	DV: Interview Government Office Region

8.2 Variable names

The variable names in the dataset are based on those used in the CAI program and are documented in the questionnaire. These variable names are prefixed with 'W8' denoting the wave/sweep of the cohort study. The remaining characters have kept as close to the questionnaire documentation as possible and therefore have not been truncated to a maximum limit.

For multi-coded variables, where a single question produces more than one response, a suffix has been used to identify the iteration. 0A, 0B, 0C, ..., AA, AB has been used to denote the 1st, 2nd, 3rd, ..., 26th, 27th iteration respectively.

Examples of multi-coded variables in the questionnaire include:

Multi-coded variables	**Overarching label**
W8SUBDEG0A – W8SUBDEGBJ	Subject of degree
W8BENO0A – W8BENOAJ	State benefit type

Derived variables in the dataset 'NS8_2015_Derived_variables' are given the prefix "W8D". These variables were constructed as part of the data preparation for archiving, while the derived variables with the prefix "W8" in the other deposited datasets were computed in the CAI program.

8.3 Variable labels

The variable labels included in the datasets give an indication of the question content. Multi-coded variables have been given a common prefix based on the question content. Variables derived in the CAI program, and those derived separately and included in the derived variables dataset have been given the prefix "DV".

8.4 Value labels

The value labels for valid responses are based on the question responses used in the CAI program as documented in the questionnaire documentation. Value labels have been individually reviewed and amended, where necessary.

Missing values for variables used in the CAI program have been consistently labelled as follows:

-9 Refused

-8 Don't know

-1 Not applicable

Missing values for derived variables have been given the label "Insufficient information" for value -8 unless otherwise stated.

8.5 Variable order

The order in which variables appear in the datasets broadly follows the order of modules and sections within modules, of the CAI program as documented in the questionnaire.

8.6 Unfolding brackets

Unfolding brackets were used on a number of questions within the Housing, and Finance modules. If respondents were unable or unwilling to give an exact amount as an answer to a particular question (for example exact income), unfolding brackets were used to gain an estimate of that amount. So for instance, if the cohort member could not answer a question on income, (s)he was asked whether his/her income was above, below or about a randomly chosen income amount from a given set of four values i.e. this is the **START** amount. (S)he could then be asked a series of similarly structured questions in order to narrow down the amount. If the answer is 'more than', the question is re-asked with the amount increased to the next in the series (unless it is already at the highest value). Similarly, if the answer is 'less than' the question is re-asked with the amount decreased to the next in the series (unless it is already at the lowest value). From the start amount, the sequence of responses is restricted to one direction until an approximate or bounded response is received. For the majority of questions, a range of different periods were covered with different unfolding brackets values according to the period, i.e. 1 week, 2 weeks, 4 weeks/calendar month, 1 year/one-off payment. Respondents who chose other periods were asked about monthly payments, yearly payments or one-off payments depending on the question.

The following example has been chosen to aid interpretation of the responses:

MOGP: Whether last total monthly mortgage instalment was greater than, less than or about the value. Range of values: (1300 800 600 450)

Ex	MOGP	MOGP _MAX	MOGP _MIN	MOGP_ START	MOGP_ CLOSE	Sequence of responses from start
1	1 (greater than min)	600	450	1300	-1	<1300, <800, <600, >450
2	1 (greater than min)	-1	1300	600	-1	>600, >800, >1300
3	2 (about)	800	-1	800	600	<800, about 600
4	2 (about)	-1	-1	450	450	about 450
5	3 (less than max)	800	600	450	-1	>450, >600,<800
6	3 (less than max)	450	-1	600	-1	<600,<450
7	-9 (Refused)	800	-1	800	600	<800, refused amount at 600

- Ex 1 MOGP=1: The final response is the minimum value shown in MOGP_MIN. This value lies within the range (600,450).
- Ex 2 MOGP=1: The final response is the minimum value shown in MOGP_MIN. This value lies outside the range of given values, that is, it is greater than 1300.
- Ex 3 MOGP=2: The final response is the amount shown in MOGP_CLOSE. This value is close to 600.
- Ex 4 MOGP=2: The final response is the amount shown in MOGP_CLOSE. This value is close to the start value 450.
- Ex 5 MOGP=3: The final response is the maximum value shown in MOGP_MAX. This value lies within the range (800,600).
- Ex 6 MOGP=3: The final response is the maximum value shown in MOGP_MAX.

This value lies outside the range of given values, that is it is less than 450.

- Ex 7 MOGP=-9: The amount answered 'Refused' is shown in MOGP_CLOSE, that is, refused to answer greater than, less than or about at value 600.

8.7 Known issues and data cleaning

Interviewers carried out data editing in the field where inconsistencies were highlighted through soft and hard checks. 'Hard' checks did not allow entries outside a given range (and had to be resolved by the interviewer at the time of the interview). 'Soft' checks required the interviewer to check and confirm what (s)he had entered. These enabled interviewers to clarify and query data discrepancies directly with the respondent during the interview.

Following checks by the fieldwork agency and additional checks at CLS, the following issues have been identified:

- CINTRO: Childcare used
  Issue: Also asked if children not living with CM (NCHPRES=3)
  Action: These responses have been set to -1 'Not applicable' in line with the original questionnaire

- NEGRCK: Check gross pay against net pay
  Issue: Incorrect operator used, specification states greater than or equal to and program is greater than
  Action: A flag W8NEGRCKF has been added to indicate whether a script error has occurred for cohort member

- W8MAKESICCODE: Type of organisation work for in main job
  Issue: 42 codes exist consisting of 4 digits instead of 5 digits.
  Action: To advise users that these codes are prefixed with 0 in the Standard Industrial Classification of Economic Activities (SIC) 2007 and belong to Sections A or B.

- Backcoded variables
  Issue: Questions that include 'Other (please specify)' categories allow the respondent to give open text responses that are back-coded after the interview is completed. Some of these variables are used in filtering cases to subsequent questions. Where back-coding has occurred after the interview, the value will not be used for filtering. In these cases the following flag variables have been added to indicate whether the expected filtering has not taken place.
  Action: W8PARTDOF has been added for W8PARTDO: Partner's main activity
  Action: W8SIBE0EF has been added for W8SIBE0E: Disability related benefit type: Disability Living Allowance, PIP
  Action: W8ACTCF has been added for W8ACTIVITYC: Main current activity

- Continuous variables
  Issue: Some continuous variables may contain outliers and/or implausible values.
  Action: None

- Unfolding brackets variables
  Issue: For a set of income and payment questions, cohort members were asked to provide a stated *amount* and a time *period*. If the response to either or both was -8 (Don't know) or -9 (Refused), they were routed to the respective unfolded brackets question. For a few questions, cohort members were not routed to the unfolded

brackets question where the time period given was ‘Other’. In these cases, flag variables have been added to indicate whether the expected filtering has not taken place.
Action: W8REG5F has been added for W8REG5, W8NAOBF has been added for W8NAOB, W8USOBF has been added for W8USOB, W8SJOBF has been added for W8SJOB, and W8PUOBF has been added for W8PUOB.

9 Documentation

In addition to this Guide, the following documentation accompanies the data deposit:

- Age 25 Survey Questionnaire
- Age 25 Survey Technical Report and Appendices
- Age 25 Survey Derived Variables Guide

10 References

Links to supporting documents

Construction of weights Wave 1 to 7. DfE webpage.

Next Steps Age 25 Survey: Technical report and appendixes. CLS website.

Next Steps Age 25 Survey: Questionnaire. CLS website.

User Guide Wave 1 to 7. DfE website.

User Guide Wave 1 to 7. UKDS website.

Alcohol Use Disorders Identification Test Consumption (AUDIT C).

References

Côté, J. E. (1997). An empirical test of the identity capital model. *Journal of Adolescence, 20*, 421-437.

Conroy, R.M. (2005). Stings in the tails: Detecting and dealing with censored data. *Stata Journal. 5:* 395-404.

Enders, C.K. (2010). *Applied missing data analysis*. Guildford Press

Golderberg, D., Williams, P. (1988). *A user's guide to the General Health questionnaire*. NFER-Nelson.

Office for National Statistics (2015). *Harmonised Concepts and Questions for Social Data Sources: Primary Principles. Long-lasting Health Conditions and Illnesses; Impairments and Disability.* Retrieved from ONS website.

Lefcourt, H. (1991) Locus of control. In J. Robinson, P. Shaver, L. Wrightsman (Eds), *Measures of Personality and Social Psychological Attitudes* (pp. 413-493). London: Academic Press.

Seaman, S., Galati, J., Jackson, D. and Carlin, J. (2013). What is meant by 'missing at random'? *Statistical Science*, *28*, 257-68.

Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res*, *22*, 278-95.

StataCorp. (2007). Stata Statistical Software: Release 10. College Station, TX: StataCorp Lp.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M. and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338.

Stewart, M.B. (1983). On least-squares estimation when the dependent variable is grouped. *Review of Economic Studies*, *50*, 737-53.

White, I.R., Roystonm P, Wood, A.M. (2011). Multiple imputation using chained equations: Issues and guidance for practice*. Stat Med, 30,* 377-99.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, *141*, 1281-1301.

Appendix A: Response and weights

Multiple imputations

We note that in Inverse Probability Weighting (IPW) analyses to account for missing data, the predictors of response are fully, or almost fully observed ("complete")(Seaman and White, 2011). This typically occurs in instances where strictly monotone missing data patterns are observed. In this instance, missing data for predictor variables were observed and were either due to non-monotone non-response (17.1%) or item missingness. Multiple Imputation with chained equations was employed using the MI command in Stata 13 resulting in a sample size of 16,122 with complete records on all predictors of response. We used 50 imputations since housing tenure in sweep 7 had the highest level of item missingness (with 50% missing).

We note that multiple imputation returns valid estimates assuming the data are Missing at Random (MAR) (Enders, 2010, Seaman et al., 2013, Sterne et al., 2009). This implies that any differences between the missing values and the observed values can be explained by the variables that were included in the imputation models. Put differently, conditional on the variables in the imputation model, missingness in not due to unobserved or observed variables not included in the model.

Table A1: Logit response model predicting response at sweep 8

Response in sweep 8	Odds Ratio	SE.	t	P>t	95% Confidence Interval	
Marital Status of main parent s1, reference: single						
Married	1.42	0.06	8.4	0	1.31	1.54
Cohabitating	1.07	0.08	0.95	0.34	0.93	1.23
Highest qualification of main parent s1, reference: degree						
HE less than degree	0.94	0.07	-0.82	0.41	0.82	1.08
GCE A level	0.92	0.06	-1.2	0.23	0.8	1.05
GCSE grades A-C	0.85	0.05	-2.61	0.01	0.75	0.96
Level 1 or below	0.71	0.06	-4.16	0	0.61	0.84
Other qual/No qual	0.75	0.05	-4.11	0	0.66	0.86
Employment Status of main parent s1, reference: doing work for more than 30 hours per week						
Doing paid work for fewer than 30 hours per week	1.09	0.05	2.08	0.04	1.01	1.18
Unemployed, other	0.92	0.04	-1.82	0.07	0.85	1.01
Family's NS-SEC class s1, reference: Higher Managerial and professional						
Lower managerial and professional	0.85	0.05	-2.68	0.01	0.75	0.96
Intermediate	0.84	0.05	-2.7	0.01	0.75	0.95
Semi-routine/Routine	0.82	0.06	-2.86	0	0.72	0.94
long term unemployed	0.78	0.07	-2.71	0.01	0.64	0.93
Sex of cohort member s1, reference: Woman						
Man	0.65	0.02	-12.69	0	0.6	0.69
Ethnic group of cohort member s1, reference: Non-White						
White	1.12	0.05	2.82	0.01	1.04	1.22
House tenure s7, reference: Rented						
Owned	1.07	0.05	1.48	0.14	0.98	1.17
Current main activity at time of interview s7, reference: Education						
Paid work	0.94	0.05	-1.2	0.23	0.85	1.04

Other	0.89	0.05	-1.96	0.05	0.79	1
Whether cohort member ever identified as having special education needs s1, reference: No						
Yes	0.93	0.04	-1.76	0.08	0.85	1.01
Whether cohort member ever tried cannabis s7, reference: No						
Yes	0.93	0.04	-1.72	0.09	0.86	1.01
Government Office Region s1, reference: North East						
North West	0.98	0.09	-0.2	0.84	0.82	1.17
Yorkshire and The Humber	0.94	0.09	-0.64	0.52	0.78	1.13
East Midlands	0.99	0.1	-0.13	0.89	0.81	1.2
West Midlands	0.92	0.09	-0.89	0.38	0.77	1.1
East of England	1.1	0.11	1	0.32	0.91	1.33
London	0.92	0.08	-0.89	0.37	0.77	1.1
South East	0.99	0.09	-0.06	0.95	0.83	1.19
South West	1.01	0.1	0.15	0.88	0.83	1.24
Month of interview s7, reference: May 2010						
Jun-10	0.84	0.04	-3.31	0	0.76	0.93
Jul-10	0.78	0.05	-3.77	0	0.68	0.89
August to October 2010	0.74	0.06	-4.04	0	0.63	0.85
Survey mode s7, reference: CAPI						
CATI	1.1	0.08	1.24	0.22	0.95	1.27
CAWI	1.39	0.11	4.28	0	1.19	1.62
Constant	1.21	0.17	1.36	0.17	0.92	1.6
N			16,122			

Note: The analytical sample in Table A1 consists of 16,122 observations.

Table A2: sweep 8 overall cross-sectional weight

Variable	N	Mean	Std. Dev.	Min	Max
Weight: All Sweep 8 respondents	7,707	1.02	0.70	0.04	6.03