Next Steps
Age 32 Sweep (Sweep 9)

User Guide (Version 1)

August 2024

CENTRE FOR LONGITUDINAL STUDIES

UKRI Economic and Social Research Council

Contact

Data queries: help@ukdataservice.ac.uk

Questions and feedback about this user guide: clsdata@ucl.ac.uk.

Authors

Tugba Adali, Sarab Rihal, Liam Wright, Richard Silverwood, Matt Brown, Morag Henderson, Aida Sanchez, Liam Curran.

How to cite this guide

Adali, T., Rihal S., Wright L., Silverwood R., Brown M., Henderson M., Sanchez A., Curran L. (2024) *Next Steps Age 32 Sweep (Sweep 9) User Guide (Version 1)*. UCL Centre for Longitudinal Studies

You should cite the data and acknowledge CLS following the guidance from https://cls.ucl.ac.uk/data-access-training/citing-our-data/

This guide was published in August 2024 by the UCL Centre for Longitudinal Studies (CLS).

Centre for Longitudinal Studies (CLS)

UCL Centre for Longitudinal Studies

UCL Social Research Institute

University College London

20 Bedford Way, London WC1H 0AL

www.cls.ucl.ac.uk

The UCL Centre for Longitudinal Studies (CLS) is an Economic and Social Research Council (ESRC) Resource Centre based at the UCL Social Research Institute, University College London. It is home to a unique series of UK national cohort studies. For more information, visit www.cls.ucl.ac.uk.

This document is available in alternative formats. Please contact the Centre for Longitudinal Studies:

Email: clsdata@uck.ac.uk

Contents

About Next Steps

Next Steps is a longitudinal cohort study, following a nationally representative group of nearly 16,000 people born in England in 1989-90. The study began when cohort members were 14 years old. With sweeps every year for the first seven years, it has captured rich information about their educational trajectories during adolescence.

Next Steps has since documented early adulthood experiences at Age 25 and this new data deposit captures how the cohort members are faring at Age 32. These adulthood sweeps have a wider scope, and include measures of health, wellbeing, family formation and labour market outcomes (among others), providing unparalleled insight into the many different aspects of this millennial generation's lives.

A vital source of evidence, Next Steps has had a major influence on national education policy and cast light on a wide range of important social issues, including the association between zero-hours contracts and mental health.

1. Introduction

The Next Steps Age 32 Sweep (Sweep 9) took place between April 2022 and September 2023. It was designed and managed by the Centre for Longitudinal Studies (CLS) at the UCL Faculty of Education and Society (IOE), and fieldwork was carried out by Ipsos. It was funded by the Economic and Social Research Council. The Age 32 Sweep is the ninth sweep of the study. The first seven sweeps of data were collected annually between 2004 and 2010, when the study was run by the Department for Education and known as the Longitudinal Study of Young People in England (LSYPE). The eighth sweep was run when cohort members were aged around 25, between 2015 and 2016, by CLS. In addition, three online[1] surveys were conducted during the COVID-19 pandemic, when data was also collected simultaneously from participants in the MRC National Survey of Health and Development, 1970 British Cohort Study, 1958 National Child Development Study, and the Millennium Cohort Study.

The Next Steps Age 32 Sweep used a sequential mixed-mode design. Participants were first invited to participate by web. After 3 weeks non-responders were then contacted by face-to-face interviewers. In addition to in-home visits, interviewers were also able to offer a range of other modes. There was also an online 'mop-up' survey where non-respondents after the face-to-face fieldwork period were re-invited to participate online, with most being invited to complete a shorter version of the questionnaire at this point.

Interviews were completed with 7,279[2] cohort members including 183 interviews completed by participants residing outside of England at the time of the survey. The response rate for the study is 53% (both when cohort members residing outside of England are considered eligible to participate, and when they are not). A full account of the study development and fieldwork procedures can be found in the Next Steps -

[1] While the first two waves were fully online, there was telephone follow up in Wave 3. See the COVID-19 Survey in Five National Longitudinal Studies Waves 1, 2 and 3 User Guide (Version 4) for details.

[2] A total of 7,284 cohort members took part in this sweep, however five cohort members requested their data to be deleted before the data were deposited, so their data was never included in the research data.

Sweep 9 Survey Technical Report produced by Ipsos, which accompanies this data deposit.

This user guide provides information about the data arising from the Next Steps Age 32 Sweep and accompanies the deposit of the data at the UK Data Service.

In addition to this user guide the Age 32 Survey data deposit includes:

Next Steps - Sweep 9 (Age 32) Questionnaire

Next Steps - Sweep 9 (Age 32) Online Short Mop-up Questionnaire

Next Steps - Sweep 9 Survey Technical Report

Next Steps Age 32 Sweep (Sweep 9) Derived Variables Guide

Data, questionnaires and user guides for all previous sweeps are also available at UKDS. All datasets use a common ID – NSID.

2. The Age 32 Sweep

The Age 32 Sweep aimed to provide data for research and policy on the lives of this generation of adults in their early 30s. Since the last sweep at Age 25, many of the cohort will have experienced important life transitions in relation to work and careers, education and training, partnerships, children and their housing situation. However, pathways to adulthood are varied and complex for this generation. For example, rather than making straightforward transitions from education to work, and to independent living and family formation, this generation of adults are now more likely to move between education and work, between living at home and independently and may delay family formation as a result. The Age 32 Sweep also collected information on many wider aspects of cohort members' lives including health and wellbeing, politics and social participation, identity and attitudes as well as capturing personality, resilience, working memory and financial literacy.

The Age 32 Sweep involved asking cohort members to 1) complete a survey 2) provide consent to give the Next Steps study team access to information held in various administrative records and 3) provide a saliva sample so that DNA could be extracted for genetic research.

The Age 32 Sweep was conducted by Ipsos and funded by the Economic and Social Research Council (ESRC).

Ethical approval was provided by Cambridge Central Research Ethics Committee (REC reference: 22/EE/0052).

The fieldwork commenced in April 2022 and was completed in September 2023.

3. Fieldwork

3.1 Issued sample

A total of 13,090 study members were initially issued to the Age 32 Sweep. The issued sample was comprised of all study members who had previously taken part in Next Steps other than 1) those who had previously indicated they did not wish to be contacted, 2) those who were known to have died, 3) those who were permanently untraced, and 4) those who were in prison or on probation.

After the mainstage fieldwork started, a further 769 cases were traced via NHS England, resulting in new addresses being obtained. These cases were issued to the Age 32 Sweep but only invited to take part online.

In total, 13,859 cases were issued to the Age 32 Sweep. This included 193 participants known to be living outside England (for whom an email address was held). These cases were only invited to take part online.

3.2 Fieldwork Period

Next Steps Age 32 Sweep data collection took place from 25th April 2022 until 24th September 2023 (See Section 6.6.1 in the Next Steps Sweep 9 Survey Technical Report).

3.3 Fieldwork stages

A sequential mixed mode design was used in which participants were first invited to take part online. After 3 weeks non-respondents were assigned to interviewers who were able to offer a face-to-face a (F2F) interview, a video interview, completion via secondary device (where the interviewer would provide a tablet to the participant on which they could self-complete the survey and then return to collect it) and in exceptional circumstances a telephone interview. The web survey also remained

open throughout the interviewer fieldwork period. There was also a final online 'mop-up' survey for remaining non-respondents at the end of fieldwork.

To make fieldwork more manageable, the issued sample was divided into four batches/waves, released to the field in sequence. The first batch was designated as a 'soft launch' to enable testing the survey processes and provide evidence of likely response. The cases issued to the soft launch were a random sample of the full initially issued sample (out of the 13,090 cases mentioned above). The soft launch included some methodological assessments to guide fieldwork practices in the remaining stages of the study (see section 4 of the Next Steps Sweep 9 Survey Technical Report). Moreover, minor changes to the questionnaire were also implemented after the soft launch. These changes are indicated in the Questionnaire documentation accompanying this deposit.

A separate web only batch was released when new addresses for previously untraced cases were obtained via a tracing exercise conducted via NHS England, which is referred to as Wave 4 in this document.

3.4 Contact Strategy

All cohort members were initially invited to complete the survey via web (and this mode of completion remained open throughout the fieldwork period). They had three weeks to complete the web survey before being issued to face-to-face interviewers.

All cohort members with a valid e-mail address were sent a pre-notification email ahead of fieldwork to let them know of the upcoming survey, with a request to update their contact details. On commencement of web fieldwork an invitation mailing was sent by post and email which provided full details about what participation would involve and included a link to the survey and login details. Over the three weeks of web fieldwork cohort members who had not participated were sent two email reminders, two text message reminders and one postal reminder (only if they did not have a valid email address). Break-off reminders, via e-mail and text message, were sent to participants who had started or partially completed the survey (see Section 7 in the Next Steps Sweep 9 Survey Technical Report).

Three weeks after the launch of the web fieldwork, non-respondents to the web survey were contacted by interviewers who would attempt contact with them either by phone, or by personal visit depending on the case (see Section 6.3.3 in Next Steps Sweep 9 Survey Technical Report).

Whenever interviewers identified that cohort members had moved from the issued address, they carried out tracing in the field, which primarily consisted of making calls to stable contacts, or asking current occupiers of the cohort member's address (see section 6.3.5 of the Next Steps Sweep 9 Survey Technical Report). Participants that interviewers were unable to trace were sent to CLS who attempted to find new contact details via office-tracing.

Following the face-to-face fieldwork, remaining non-respondents were invited to take part in an online mop-up, which aimed to provide a final chance for cohort members to take part. This was carried out in December 2022 - January 2023 for the soft launch sample. An experiment was conducted whereby half of non-respondents were invited to complete an abbreviated version of the questionnaire and the remainder were invited to complete the full survey. The shorter survey was more successful at boosting response and so this version was used in the mop-up survey for all remaining waves which was conducted in September 2023. The variable W9FULLINT specifies whether the full/long version or the short version of the questionnaire was completed.

3.5 Pilot

Main stage fieldwork was preceded by a pilot study which was conducted between May and August 2021. At the time of the pilot it was uncertain whether home visits would be feasible for the Age 32 Sweep given COVID-19 restrictions in place at the time. The pilot therefore focused on testing the feasibility of conducting the survey using a range of alternative modes. The pilot was not conducted with Next Steps study members and as such the data collected is not included in the deposited data. For a full description of the pilot please see the Next Steps Sweep 9 Survey Technical Report (Section 2).

3.6 Incentives

In Waves 1 through 3, an incentive of £30 was offered to participants conditional on completing the survey online within the first three weeks, which decreased to £20 afterwards. This incentive approach was adopted after an experiment during the soft launch, which tested whether targeting Age 25 Sweep non-respondents with a higher amount (£35 in the first 3 weeks, £25 afterwards) could improve response amongst this group, and if a lower amount for Age 25 Survey respondents could offset this additional cost without affecting response (£25 in the first 3 weeks, £15 afterwards). The results of this experiment did not show any gains in response among Age 25 Sweep non-respondents, and hence a flat incentive was used in the remaining waves (see Section 4.1 of the Next Steps Sweep 9 Survey Technical Report). During the online mop-up phase participants were offered the same incentive that they were offered for completion after 3 weeks (i.e. £20 in most cases). During the soft-launch participants were offered a £5 incentive for providing a saliva sample. This was increased to £10 after the soft launch in order to attempt to boost saliva return rates.

4. Response at Sweep 9

4.1 Overall response and outcome codes

A total of 13,859 cohort members were issued for fieldwork. A total of 7,279 interviews were completed.

The issued sample included those who were known to be living outside of England if an email address was held. This group were only invited to participate online. Those living outside of England are not formally considered part of the target population for Next Steps and so we present two versions of the response rate, one which defines all those living outside of England as ineligible (Response Rate A) and a second which treats those living outside of England as eligible (Response Rate B). Those who were found to have died or to be in prison or on probation are treated as ineligible in both response rate calculations.

Response Rate A – Emigrants ineligible – 7,096 interviews achieved from an eligible sample of 13,354 giving a response rate of 53%.

Response Rate B – Emigrants eligible – 7,279 interviews achieved from an eligible sample of 13,820 giving a response rate of 53%.

Among the total number of 7,279 interviews, 6,943 were fully completed (95%), and 336 were partially completed (defined as having at least completed the household relationships module). The figure of 7,279 is the basis for the rest of the tables in this section.

Table 1 below shows the overall response for the cases issued to fieldwork. It shows that the main reason for non-response was not being able to trace participants (21% of the total issued sample).

Table 1. Sweep 9 overall response for issued sample

Outcome	Frequency	Percent
Productive	7,279	52.5%
Refusal	2,286	16.5%
Non-contact	958	6.9%
Untraced	2,239	16.2%
Other unproductive	1,058	7.6%
Ineligible (deceased or in prison only, excluding emigrants)	39	0.3%
Total issued sample	13,859	100.0%

Note: Emigrant study members who took part are considered productive in this table, and those who did not as 'other unproductive'. Wave 4 (online only) non-respondents are classified as 'non-contact'.

4.2. Response by fieldwork stage

Web-only stage

4,721 interviews (including partials) were achieved during the three-week web phase (a 34% response rate[3]).

Face-to-face stage

8,311 participants were issued to interviewers and 2,070 interviews were achieved giving a 25% response rate for the face-to-face stage. The face-to-face stage increased the overall response rate to 49%. 50% of the interviews achieved during this phase were completed online, 35% were completed in-person and the remainder were completed by secondary device, video or telephone.

[3] Response Rate B – i.e. emigrants treated as eligible.

Online mop-up stage

6,224 non-respondents were issued to the online mop-up stage and 488 interviews were achieved giving an 8% response rate. The online mop-up stage increased the overall response rate to 53%.

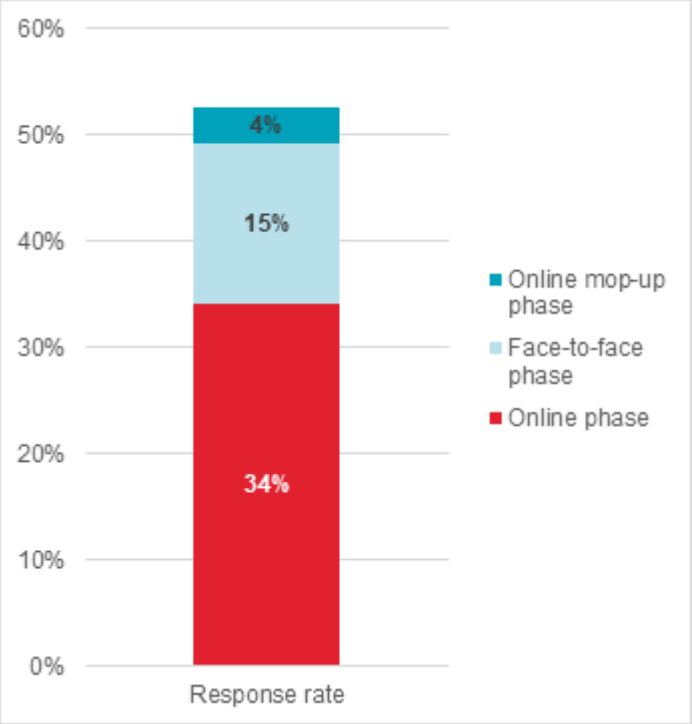


Figure 1. The contribution of the different phases of fieldwork to the overall response rate (Response rate B)

4.3. Mode of completion

Overall, Web was the most common mode for taking part at the Age 32 Sweep, with 86% of all respondents having responded via this mode. This was also the case during the face-to-face phase, where Web completions accounted for 50% of completions. Table 2 below shows the number of interviews by mode. Survey mode is denoted by the W9MODE variable in the survey data set.

Table 2. Mode of response

Mode	Frequency	Percent
Web	6,239	86%
In-home	721	10%
Secondary device	171	2%
Video	9	<1%
Telephone	139	2%
Total	7,279	100%

4.4. Response by prior sweep participation

Table 3 below shows that over 75% of productive cases were last interviewed in Sweep 8. The remainder of cases were last interviewed at earlier sweeps.

Table 3. Distribution of Age 32 Sweep respondents by prior wave participation

Sweep	Frequency	Percent
Sweep 1	166	2.3%
Sweep 2	90	1.2%
Sweep 3	88	1.2%
Sweep 4	94	1.3%
Sweep 5	121	1.7%
Sweep 6	146	2.0%
Sweep 7	998	13.7%
Sweep 8	5,576	76.6%
Total	7,279	100.0%

A thorough review of response patterns in Next Steps to date is included in Chapter 7.

5. Questionnaire

5.1 Overview

The Age 32 Sweep was comprised of 11 modules which covered household relationships, housing, activities and employment, finance, education, health, identity, self-completion (computer-assisted self-interviews - CASI), data linkage, saliva sample collection consent, and contact information.

The CASI module consisted of relatively more sensitive questions, including well-being, depression, drinking, smoking, drugs, crime, gender identity, sexual behaviour, pregnancy history, and childhood circumstances. During in-home and video interviews, interviewers were asked to ensure cohort members complete this section on their own, however there could exceptions if the cohort member required assistance.

Table 4 below summarizes how the CASI module worked in different modes:

Table 4. Self-completion section in different survey modes

Mode	Self-completion administration
Online	The entire questionnaire was self-completed including the CASI module.
In-home interview	The CAPI (computer-assisted personal interviewing) tablet was handed over to the cohort member to complete. Cohort members could request the interviewer to read-out the self-completion module should they experience difficulty with reading or sight problems. Cohort members also had the option to refuse to complete the entire section. The variable W9CASISTART denotes whether this module was completed by the cohort member, with the help of the interviewer, or whether was refused completely.
Remote method: Secondary device	The entire questionnaire was self-completed including the CASI module.

Remote method: MS Teams interview	The script functioned as a normal online survey. Cohort members were provided with an online link to complete the self-complete module in a separate browser window on their own computer where the interviewer could not observe the participant's responses. As in CAPI interviews, the variable W9CASISTART denotes whether this module was completed by the cohort member, with the help of the interviewer, or whether was refused completely. For cohort members that requested assistance, the interviewer would offer to share their screen and could select one of the following options: • The cohort member could read the content on the screen and instruct the interviewer which number to capture. • The interviewer read the question to the cohort member in full. This nature of the assistance as above is provided in the SELFCOMP_CASIINTCOM variable.
Remote method: Telephone with e-showcards	The self-completion module was not treated separately from other modules, interviewer was requested to read out the self-completion module to the cohort member.

A summary of the content of each module is provided below in Table 5.

The majority of the those who took part during the final online mop-up phase completed an abbreviated version of the questionnaire. The abbreviated questionnaire was designed to include the core measures which would be used most widely by researchers. Table 5 shows which content was included in the abbreviated mop-up questionnaire. As mentioned earlier, users can see which cases completed the abbreviated version of the questionnaire via the W9FULLINT variable.

Table 5. Questionnaire content at Sweep 9

Module number	Module title	Full questionnaire content	Whether included in the 20-minute online mop-up questionnaire
1	Household relationship	Introductions	Yes
		Confirmation of cohort member details: • Name, DOB and contact information (address, telephone number(s), email	Yes
		Partner grid: • Confirmation of partner details at last sweep (if interviewed in the last sweep): name, sex, gender, DOB and relationship to CM • Co-habiting relationships – since previous interview (if taken part after September 2006) or since September 2006 (if not taken part after September 2006) – includes start and end dates of living together, name of partner, sex and gender, and DOB/age • Current marital status – including dates of marriages/divorces/deaths • (Current) Non-cohabiting partner: name, sex, gender, DOB, start date of relationship, previous cohabitation and end date of living together	Yes
		Child grid: • Confirmation of children recorded at last sweep: name, DOB, sex, gender (if 11 or more years), relationship to cohort member	Yes

Module number	Module title	Full questionnaire content	Whether included in the 20-minute online mop-up questionnaire
		• Details of any other (new) children CM considers to be a parent of: name, DOB, sex, gender (if 11 or more years), relationship to cohort member • Whether child is current partner's biological child • Whether child is currently living, has always lived since birth, or ever lived with CM • When child last lived with CM (age or DOB) • Date child started living with CM (if not living with them at last interview) • Date of child's death (if applicable) • Non-resident children: who child lives with, frequency of in-person contacts, frequency of overnight stays, child's maintenance • Non-resident parents: frequency of in-person contacts with child, frequency of overnight stays, child's maintenance • Childcare: Childcare used for children under 16	
		Other household members: • Confirmation of details from last sweep: name, DOB/age, sex, gender (if 11 or more years), relationship to cohort member (if still living with this person); end dates of living together (if no longer living with this person)	Yes

Module number	Module title	Full questionnaire content	Whether included in the 20-minute online mop-up questionnaire
		• Details of any other (new) household members not recorded: name, DOB/age, sex, gender (if 11 or more years), relationship to cohort member, start dates of living with this person	
2	Housing	Current housing: • Whether living at same address as last sweep (if interviewed in the last sweep) • Dates of moving at current address • Type of accommodation – tenure, ownership details, tenancy • Room number • Property purchase price, source of funding, dates of purchase • Whether current property is first owned or whether ever owned (or shared ownership of) a property, age first owned a property • Rent payments, rent deductions • Outstanding mortgage payments • Satisfaction with accommodation • Ownership of other property, value of property/outstanding mortgage payments	Yes, but only included: • Whether living at same address as last sweep (if interviewed in the last sweep) • Dates of moving at current address • Type of accommodation – tenure, ownership details, tenancy

Module number	Module title	Full questionnaire content	Whether included in the 20-minute online mop-up questionnaire
		Previous housing: • Age first moved out of parents' home • Number of addresses lived at outside of parents' home	No
		Homelessness after 16: • Whether have ever or number of times have been homeless after age 16 • Age when (first) became homeless • (Total) Period of homelessness • Whether homeless on their own or with family (last period) • Reason(s) to become homeless (last period) • Where stayed while homeless (last period)	No
3	Activities and employment	Economic Activity History - economic activity since last interview (if taken part after September 2006)/September 2006 (if not taken part after September 2006): • Economic activity, full -time or part -time - if in work or education, institution of study - if in education, temporarily or long -term – if sick or disabled, date started/ended economic activity, Reasons for change in activity • Employment at 25 (if not taken part in last interview): job title, job description, main product of organisation, type of organisation	No

Module number	Module title	Full questionnaire content	Whether included in the 20-minute online mop-up questionnaire
		Current employment: • Whether full-time/part time employed • Job details – job title, job description, main product of organisation, supervising responsibilities, size of workplace/no of employees, qualifications and training • Working hours • Shift work and work at night between 10pm and 7am • Work security • Satisfaction with current job • Stress at work	Yes, but only included: • Whether full-time/part time employed • Job details – job title, job description, main product of organisation, supervising responsibilities, size of workplace/no of employees, qualifications and training • Working hours • Satisfaction with current job • Stress at work
		Second job: • Hours worked	No
		Prospective employment • Reasons for unemployment	No

Module number	Module title	Full questionnaire content	Whether included in the 20-minute online mop-up questionnaire
		• Whether looking for employment • Methods used in search for employment • • Reasons why not looking for employment	
		Employment details for first job after full time education: • Details of first – job title, job description, main product of organisation	No
		Employment support: • How found out about current/last job • Whether needed highest qualification to get current job/last job	No
		Partner current activity: • Whether full-time/part time employed • Whether in full-time/part time education • Temporary or long-term sick/disabled • Job details –including job title, job description, main product of organisation, supervising responsibilities, qualifications and training, size of workplace/no of employees	Yes
		Economic shocks experienced since coronavirus: • Whether experienced by cohort member or their (cohabiting) partner • Time period of economic shock	No
4	Finance	Current pay/salary:	No

Module number	Module title	Full questionnaire content	Whether included in the 20-minute online mop-up questionnaire
		• Gross and net pay • Self -employed income • Take home income second job • Other income from paid work	
		(Cohabiting) Partner pay/salary: • Net pay / Self -employed income	No
		Benefits (including Coronavirus state benefits): • Universal credit • Types/amounts of benefits received	No
		Other income: • Source of income, total amount received in last month • Total (household) income (cohort member and cohabiting partner) after tax and period it covered	Yes, but only included: • Total (household) income (cohort member and cohabiting partner) after tax and period it covered
		Pensions: • Whether member of pension scheme, type of pension scheme • Whether currently contributing to pension	No

Module number	Module title	Full questionnaire content	Whether included in the 20-minute online mop-up questionnaire
		• Expected retirement income sources	
		Debt: • Types of Debt, Total amount owed • Self-assessed management of own finances • Whether has difficulty paying bills • Financial position during Covid	No
		Savings and Investments: • Accounts held • Amount of savings and investments	No
5	Education	Academic and vocational qualifications gained and currently studied • Details of undergraduate degree: university from which obtained degree/or currently studying, whether degree single or joint honours, degree classification, year of graduation, subject of degree , whether university was first choice • Details of first choice university: whether applied for same subject at first choice university, single/joint honours at first choice university, subject of degree applied for at first choice university • Details of post -graduate degree: university from which obtained/or currently studying Master's degree, masters subject, university from	Yes, but with reduced detail: • Highest academic/vocational qualifications • Current academic/vocational qualifications studied for

Module number	Module title	Full questionnaire content	Whether included in the 20-minute online mop-up questionnaire
		which obtained/or currently studying Doctorate degree, subject of doctorate degree	
		Fees paid and funding received for UG degree • Whether received a loan by the Student Loans Company, whether making student loan repayments • Scholarships, grants and bursaries received • How paid for fees and living expenses while at university/college	No
		Partner's academic and vocational qualifications	No
6	Health	General health • Self-rated general health • ONS harmonised questions on long-standing illness/disability	Yes, but with reduced detail on long-standing illness/disability.
		COVID-19: • Experience of COVID-19 • Tests taken and results • Long COVID • Vaccinations	Yes, but with reduced detail.
		Height and weight: Self-reported height and weight	Yes
		Exercise	No

Module number	Module title	Full questionnaire content	Whether included in the 20-minute online mop-up questionnaire
		Sleep Fruit and fizzy drinks intake	
7	Identity	Ethnicity: • Ethnicity • Partner ethnicity • Self-rated importance of ethnicity	Yes, but only included: • Ethnicity
		National identity European identity Whether born in the UK /Country of birth and Year moved to the UK Where mother/father was born in UK / Country of birth	Yes, but only included: • whether born in the UK.
		Social support: • Number of close friends, current relationships with friends, family members, community members	No
		Attitudes: • Right to abortion, importance of women at work, environment issues	No
		Civic engagement and politics: • Attendance of meeting meetings for local groups or voluntary organisations • Unpaid voluntary work	No

Module number	Module title	Full questionnaire content	Whether included in the 20-minute online mop-up questionnaire
		• Interest in politics • Participation in protests • Voting in Dec 19 general election and the EU Referendum in 2016	
		Trust of other people	No
8	CASI (Self Completion Module)	Working memory and concentration • Digit Span Task (Backwards)	No
		Financial literacy: • Inflation and interest	No
		Opinion on immigration	No
		Personality traits • BIG5 Personality Traits (extroversion, agreeableness, openness, conscientiousness, and neuroticism) • Short GRIT scale	Yes, but only included: • BIG5 Personality Traits (extroversion, agreeableness, openness, conscientiousness, and neuroticism)
		Well-being and life satisfaction	No

Module number	Module title	Full questionnaire content	Whether included in the 20-minute online mop-up questionnaire
		• ONS Personal wellbeing (life satisfaction, worthwhile, happiness, anxiety) • Overall life satisfaction	
		Mental health and well-being • General Health Questionnaire (GHQ12) • Generalized Anxiety Disorder questions (GAD2) • Patient Health Questionnaire items (PHQ2) • Diagnoses depression/serious anxiety and treatment	Yes, but only included: • General Health Questionnaire (GHQ12) • Generalized Anxiety Disorder questions (GAD2) • Patient Health Questionnaire items (PHQ2)
		Loneliness: • UCLA loneliness scale	No
		Alcohol and smoking: • Alcohol Use Disorders Identification Test (AUDIT) • Smoking – whether smokes regularly, number of cigarettes usually smoked a day, number of cigarettes usually smoked a day, age when started / last smoked cigarettes regularly	No

Module number	Module title	Full questionnaire content	Whether included in the 20-minute online mop-up questionnaire
		• Use of electronic cigarette/vaping device	
		Illegal drugs • Drug use – ever, last 12 or 4 months	No
		Contact with the Criminal Justice System in the last 12 months	No
		Sexual orientation, gender identity and sexual behaviour	No
		Smear testing and menarche	No
		Pregnancy histories, fertility treatments and family planning	No
		Relationship quality and domestic violence	No
		Workplace abuse and discrimination	No
		Adverse childhood experiences • Childhood health • Difficult events in childhood (parental separation, death, violence) • Financial difficulties in childhood	No
9	Data linkage	Cohort member permission (if not given at Age 25 Sweep) to add: • Health records (NHS Digital) • Education records (DfE/HESA, UCAS, SLC) • Economic records (HMRC, DWP) • NI number (if permission to add SLC, HMRC or DWP records given)	No

Module number	Module title	Full questionnaire content	Whether included in the 20-minute online mop-up questionnaire
		• Criminal records (MOJ)	
10	Saliva consent	Permission to send saliva kit post-interview	No
11	Contact information	Work contact details (telephone number) (Cohabiting) Partner contact details (telephone number, email) Stable contact details (address, telephone number, email) Social media (Facebook, Twitter, Instagram) details	Yes

Note: A summary of the questionnaire content at the first seven sweeps of the study is available in the LSYPE User Guide to the Datasets: Wave 1 to Wave 7, and for the eighth sweep of the study in the Next Steps Sweep 8- Age 25 Survey User Guide at the UKDS website.

Across all modes interviews took, on average, 59 minutes to complete. Completing via web took 54 minutes on average whilst participating face-to-face took 86 minutes on average.

The questionnaire was designed to minimise the risk that mode of completion would affect comparability of the data collected in the different modes.

Variations to questions 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 web self-completion.

Section 8 of this User Guide discusses how researchers should take mode effects into account in analysis.

The questionnaire was scripted and implemented by Ipsos. It was extensively tested both by Ipsos, and CLS.

5.2 Data linkage

During the Age 25 Sweep participants were asked to provide consent to link survey data with data held by various government departments and agencies. In the Age 32 Sweep those who had not previously consented (either because they did not take part or because they chose not to consent at 25) were asked for their consent (again).

Consents covered linkage to the following records:

• 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 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)

• Police National Computer (PNC) records covering arrests, cautions and sentences, held by the Ministry of Justice

A full description of the consent process and consent rates obtained is provided in the Next Steps Sweep 9 Survey Technical Report (Section 5.4.1 and Section 8.7).

For up to date information about the availability of linked data for research visit: https://cls.ucl.ac.uk/data-access-training/linked-data/

5.3 Other special features

Saliva sample collection

At the end of the Age 32 Sweep cohort members were asked if they would be willing to provide a saliva sample from which DNA would be extracted for genetic research. This was the first time that Next Steps participants have been asked to provide any biological sample. Please see the Next Steps Sweep 9 Survey Technical Report for details on how these samples were collected (section 5.4.4). Consent and return rates are shown in Table 6.

Table 6. Saliva sample consents and return rates

Eligible for saliva consent	Saliva consent given		Sample received		
n	**n**	**Out of eligible (%)**	**n**	**Out of consents (%)**	**Out of eligible (%)**
6,352	3,591	57%	1,733	48%	27%

In total, 57% of study members who were asked to provide a saliva sample consented to providing sample (which meant they were provided with a kit to do so). In the event 48% of those who consented, and 27% of those who were eligible sent a sample back to the lab.

DNA was extracted from the samples and subsequently genotyped. Following quality control, genotype data is available for 1,568 study members. Please see details here: https://cls-genetics.github.io/docs/Next_steps.html

This data is available for research via application to the CLS Data Access Committee: https://cls.ucl.ac.uk/data-access-training/data-access/accessing-data-directly-from-cls/

Occupation coding

Participants were asked to provide details about their current job, their first job (if not previously collected at Age 25) and the job they were doing at Age 25 (if not previously collected at Age 25). Those with a cohabiting partner were also asked to provide details about their partner's job.

All occupations were coded to the four-digit standard occupation coding frame (SOC 2020).

Occupation coding was conducted during the interview via a look-up system. Participants (or interviewers) entered their job title and keywords which described their occupation and were then presented with a list of potential occupations and accompanying SOC codes from which they were asked to select the most

appropriate. Participants could amend the entered details if no appropriate option was displayed. Where participants were unable to select an appropriate code they were asked to provide an open-text description of their occupation which was then manually coded to SOC2020 by office coders.

The use of a look-up approach to conduct occupation coding is novel – though a similar approach was used in the Age 25 Sweep. In order to evaluate the effectiveness of this approach those who successfully selected a SOC code for their current job were also asked for an open text description which was manually coded by two independent coders.

Further details are provided in the Next Steps Sweep 9 Survey Technical Report (section 5.4.2).

The SOC2020 codes deposited in the data are those selected from the look-up by participants/interviewers during the interview – unless no code was selected, in which case the code allocated by the first office coder has been provided. NS-SEC has also been derived from the deposited SOC code.

To minimise disclosure risk, 3 digit SOC codes are included in EUL deposit. The 4 digit SOC codes are available under Secure Access (see Section 6.2).

The additional office-based codes are for research via application to the CLS Data Access Committee: https://cls.ucl.ac.uk/data-access-training/data-access/accessing-data-directly-from-cls/.

The below table summarises occupation coding variables in the EUL deposit data:

	First job	Job at Age 25	Current job	Partner's job
3 digit SOC code (per SOC2020)	W9FIRSTJOB2020	W9AGE25JOB2020	W9CURRJOB2020	W9PARTNERJOB2020
3 digit SOC code (per SOC2010)	W9FIRSTJOB2010	W9AGE25JOB2010	W9CURRJOB2010	W9PARTNERJOB2010
NS-SEC (SOC 2020)	NA	NA	W9NSSEC	NA
NS-SEC 8 (SOC 2020)	NA	NA	W9NSSEC8	NA
NS-SEC 5 (SOC 2020)	NA	NA	W9NSSEC5	NA

Cognitive assessment

In the Age 32 Sweep, for the first time study members were asked to complete a cognitive assessment task (Harvard's TestMyBrain (TMB) Backward Digit Span), a measure of working memory (Singh et al. 2021). The task involved the participant memorising a sequence of digits and then recalling them in the reverse order (for example, '1', '2' would be recalled as '2', '1'). The sequence started with two numbers and increased by one digit at a time as it progresses, reaching a maximum of 11 digits. There were two sequences at each length. Study members started with 'practice' trials and then moved on to 'test' trials for the actual exercise. The task ended when a study member failed to correctly recall either of the two sequences of a particular digit length, or correctly entered all displayed numbers up to the maximum number of digits. They also had an option to close the exercise before it was complete if they wished to do so.

The variables deposited in the NS9_2022_Main_Interview dataset are as follows: W9SCORE is the aggregate score (highest number of digits recalled correctly) for cohort members who completed the 'test' trials and range between 0 and 11. W9NUMCORRECT is the total number of correctly recalled sequences (excluding practice trials) and ranges between 0 and 20.

W9COGDEVICE shows the type of action used by the study member on their response to the first 'test' trial, required by their device (keyboard, pen/stylus, touch).

Variable name	**Variable label**
W9SCORE	Backward Digit Span: Score
W9NUMCORRECT	Backward Digit Span: Total number correctly recalled sequences
W9COGDEVICE	Backward Digit Span: Entry method (1st test trial)

A summary of the scoring (W9SCORE) is given below:

Score	Explanation
1	Failed both attempts when sequence length of response is 2
2	Failed both attempts when sequence length of response is 3
3	Failed both attempts when sequence length of response is 4
4	Failed both attempts when sequence length of response is 5
5	Failed both attempts when sequence length of response is 6
6	Failed both attempts when sequence length of response is 7
7	Failed both attempts when sequence length of response is 8
8	Failed both attempts when sequence length of response is 9
9	Failed both attempts when sequence length of response is 10
10	Failed both attempts when sequence length of response is 11
11	Successfully completed either attempt when sequence length of response is 11

Trial by trial data is available in a separate dataset NS9_2022_Cognitive_Tests containing responses for each practice and test trial taken. Reaction time is available as well as duration of time taken across the tests. This dataset covers ‘practice’ trials as well as ‘test’ trials. An additional variable W9STATE provides the input action study members used for each trial (including both ‘test’ and ‘practice’ trials), with the same response options as W9COGDEVICE.

5.4 Scales

The Next Steps Age 32 Sweep included several established scales which are listed below. Overall scores for each scale have been derived and included within the data deposit (data set name NS9_2022_Derived_variables), and are covered in this section (variables whose labels have a ‘DV:’ prefix). Further details regarding the

derivation of the scores can be found in Derived Variables Guide, and original wording used in the scales can be found in the Next Steps Sweep 9 (Age 32) Questionnaire.

5.4.1 Health module: ONS long lasting health conditions and illnesses: Impairments and Disability (ONS, 2015)

The Age 32 Sweep 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 (W9DDISEA) and whether they have a long-standing illness or condition using the European Union's Statistics on Income and Living Conditions (EU-SILC) definition (W9DDISEU) (ONS, 2015). W9DDISEA identifies individuals as disabled or not, W9DDISEU 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**
W9LOIL	Has longstanding illness
W9LOLM	Reduced day-to-day activities as result of longstanding illness
W9LOLP	Length of time day-to-day activities affected by longstanding illness
W9DDISEA	DV: Disability classification Equality act (2010)
W9DDISEU	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 (W9LOIL) and have a reduced ability to carry out day-to-day activities as a result of their illness (W9LOLM).

According to the EU-SILC definition, a cohort member is considered to be disabled if they report a longstanding illness (W9LOIL), have a reduced ability to carry out day-to-day activities as a result of their illness (W9LOLM), and this reduced ability has

lasted for more than 6 months (W9LOLP). This variable also distinguishes between those that are disabled to some extent, and those that are severely hampered (from W9LOLM).

Modified versions of the above items have been asked to cohort members in Waves 4, 6, 7 and also 8. Parents' reports have been collected in Waves 1 and 2.

5.4.2 Identity module: Social provisions

Cutrona CE, Russell DW. The provisions of social support and adaptation to stress. Advance in Personal Relationships. 1987;1:37–67

Three items were included from the 10-item Social Provisions Scale (Cutrona 1987). The Social Provisions Scale measures the availability of social support.

Cohort members were asked to think about their current relationships with friends, family members, community members and so on. They were asked to indicate the extent to which each statement described their current relationship with other people from the following responses:

1. Very true
2. Partly true
3. Not true at all

Variable Name	**Variable label**
W9SOCPROVA	Feels safe secure and happy
W9SOCPROVB	Someone to trust
W9SOCPROVC	Someone to feel close to

Above items have also been asked to Next Steps cohort members in the COVID-19 surveys.

5.4.3 Self-completion module: General Health Questionnaire (GHQ-12) (Goldberg & Williams, 1988)

Goldberg D, Williams P. A user's guide to the general health questionnaire. London: Nfer-Nelson; 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. A higher score on this scale indicates a greater likelihood of mental ill health.

Variable name	**Variable label**
W9GHQ12_1	GHQ12: Concentrate on what doing
W9GHQ12_2	GHQ12: Lost sleep over worry
W9GHQ12_3	GHQ12: Playing a useful part in things
W9GHQ12_4	GHQ12: Capable of making decisions
W9GHQ12_5	GHQ12: Felt constantly under strain
W9GHQ12_6	GHQ12: Felt couldn't overcome difficulties
W9GHQ12_7	GHQ12: Enjoy day to day activities
W9GHQ12_8	GHQ12: Face up to problems
W9GHQ12_9	GHQ12: Felt unhappy or depressed
W9GHQ12_10	GHQ12: Losing confidence in self

W9GHQ12_11	GHQ12: Thinking of self as worthless
W9GHQ12_12	GHQ12: Felt reasonably happy
W9DGHQSC	DV: General Health Questionnaire (GHQ12) score (Goldberg & Williams,1988)

The 12 GHQ items have also been asked at Sweeps 4,2, 8, as well as the COVID-19 surveys.

5.4.4 Self-completion module: GAD2 (Generalised Anxiety Disorder 2-item)

Kroenke K, Spitzer RL, Williams JB, Monahan PO, Löwe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann Intern Med. 2007;146:317-25.

The GAD-2 was based on the GAD-7, which was developed by Drs. Robert L. Spitzer, Janet B.W. Williams, Kurt Kroenke and colleagues, with an educational grant from Pfizer Inc. No permission required to reproduce, translate, display or distribute.

The Generalized Anxiety Disorder 2-item (GAD-2) is a brief initial screening tool for generalized anxiety disorder.

Respondents are asked how often they have been bothered by problems over the last 2 weeks: a) “Feeling nervous, anxious or on edge”; and b) “Not being able to stop or control worrying”, with the following response options:

1. Not at all
2. Several days
3. More than half the days
4. Nearly every day

The GAD-2 score is obtained by adding the score for each question (Total points). The score for each question is:

0 = Not at all

1 = Several days

2 = More than half the days

3 = Nearly every day

Variable name	Variable label
W9GAD2PHQ2AD0a	Feeling nervous, anxious or on edge
W9GAD2PHQ2AD0b	Not being able to stop or control worrying
W9DGAD2	DV: Generalised Anxiety Disorder 2-item

Above items have also been asked to Next Steps cohort members in the COVID-19 surveys.

5.4.5 Self-completion module: PHQ2 (Patient Health Questionnaire 2-item)

Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: Validity of a Two-Item Depression Screener. Medical Care. 2003;41:1284-92.

The PHQ-2 enquires about the frequency of depressed mood and anhedonia over the past two weeks. The PHQ-2 includes the first two items of the PHQ-9 Respondents are asked how often they have been bothered by problems over the last 2 weeks: c) “Little interest or pleasure in doing things”; and d) “feeling down, depressed or hopeless”, with the following response options:

1 = Not at all

2 = Several days

3 = More than half the days

4 = Nearly every day

The PHQ-2 score is obtained by adding the score for each question (Total points). The score for each question is:

0 = Not at all

1 = Several days

2 = More than half the days

3 = Nearly every day

Variable name	Variable label
W9GAD2PHQ2AD0c	Little interest or pleasure in doing things
W9GAD2PHQ2AD0d	Feeling down, depressed or hopeless
W9DPHQ2	DV: Patient Health Questionnaire 2-item

Above items have also been asked to Next Steps cohort members in the COVID-19 surveys.

5.4.6 Self-completion module: UCLA loneliness 3 item

Daniel W. Russell (1996) UCLA Loneliness Scale (Version 3): Reliability, Validity, and Factor Structure, Journal of Personality Assessment, 66:1, 20-40, DOI: 10.1207/s15327752jpa6601_2

Hughes, M. E., Waite, L. J., Hawkley, L. C., & Cacioppo, J. T. (2004). A Short Scale for Measuring Loneliness in Large Surveys: Results From Two Population-Based Studies. *Research on aging*, *26*(6), 655–672. https://doi.org/10.1177/0164027504268574

Three items from the 20-item UCLA loneliness scale were asked of all cohort members. They were asked to give the frequency in response to questions about current loneliness and related emotional states from the following response options:

1. Hardly ever
2. Some of the time
3. Often

A combined score is obtained by adding the score for each question (Total points). The score for each question is:

1 = Hardly ever
2 = Some of the time
3 = Often

In addition, a fourth item (How often do you feel lonely?) was included which is not part of the UCLA scale, but has been used in the Next Steps data due to space constraints.

Variable Name	Variable label
W9LONELA	How often feels lack companionship
W9LONELB	How often feels left out
W9LONELC	How often feels isolated from others
W9DLONELINESS	DV: UCLA loneliness 3 item

Above items have also been asked to Next Steps cohort members in the COVID-19 surveys.

5.4.7 Self-completion module: GRIT-8

Duckworth, A. L., & Quinn, P. D. (2009). Development and Validation of the Short Grit Scale (Grit–S). Journal of Personality Assessment, 91(2), 166–174. https://doi.org/10.1080/00223890802634290

Duckworth, A.L., Peterson, C., Matthews, M.D., & Kelly, D.R. (2007). Grit: Perseverance and passion for long-term goals. Journal of Personality and Social Psychology, 9, 1087-1101.

The grit scale was developed to measure perseverance and passion for long-term goals, originally as a 12 item scale. An 8-item version was developed later (cited above), which was used in the Age 32 Sweep.

The scale consists of 4 positively and 4 negatively worded phrases, and are answered on a 5 point response options as below:

1. Very much like me
2. Mostly like me
3. Somewhat like me
4. Not much like me
5. Not like me

For each positively worded question, a decreasing score was assigned from “Very much like me” as 5, to “Not like me” as 1. For each negatively worded question, a score was assigned from “Very much like me” as 1, to “Not like me” as 5. These were summed across all 8 questions and divided by 8. The maximum score is 5 (extremely gritty), and the lowest score is 1 (not at all gritty).

Variable Name	**Variable label**
W9GRIT180A	GRIT 1-8 - New ideas and projects sometimes distract from previous ones.
W9GRIT180B	GRIT 1-8 - Setbacks don’t discourage. Don’t give up easily.
W9GRIT180C	GRIT 1-8 - Obsessed with a certain idea or project for a short time but later lost interest.
W9GRIT180D	GRIT 1-8 - Hard worker.
W9GRIT180E	GRIT 1-8 - Often set a goal but later choose to pursue a different one.
W9GRIT180F	GRIT 1-8 - Have difficulty maintaining focus on projects that take more than a few months to complete.
W9GRIT180G	GRIT 1-8 - Finish whatever I begin.
W9GRIT180H	GRIT 1-8 - Diligent. Never give up.
W9DGRIT18	DV: Short GRIT scale

5.4.8 Self-completion module: Big Five personality traits

Lang, F. R., John, D., Ludtke, O., Schupp, J., & Wagner, G. G. (2011). Short assessment of the Big Five: robust across survey methods except telephone interviewing. Behavior Research Methods, 43, 548-567.

The Big Five personality traits, also known as the five factor model (FFM), is a model based on common language descriptors of personality. The five factors have been defined as openness to experience, conscientiousness, extraversion, agreeableness and neuroticism, often listed under the acronyms OCEAN or CANOE.

Cohort members were each asked to rate how much each of the following 15 statements applied to them using a scale of 1 to 7, where 1 is ‘does not apply to me at all’ and 7 is ‘applies to me perfectly’.

Variable Name	Variable label
W9BIGAO0A	BIG5 Personality traits BIGA-BIGO - is sometimes rude to others
W9BIGAO0B	BIG5 Personality traits BIGA-BIGO - does a thorough job
W9BIGAO0C	BIG5 Personality traits BIGA-BIGO - is talkative
W9BIGAO0D	BIG5 Personality traits BIGA-BIGO - worries a lot
W9BIGAO0E	BIG5 Personality traits BIGA-BIGO - is original, comes up with new ideas
W9BIGAO0F	BIG5 Personality traits BIGA-BIGO - has a forgiving nature
W9BIGAO0G	BIG5 Personality traits BIGA-BIGO - tends to be lazy
W9BIGAO0H	BIG5 Personality traits BIGA-BIGO - is outgoing, sociable
W9BIGAO0I	BIG5 Personality traits BIGA-BIGO - gets nervous easily
W9BIGAO0J	BIG5 Personality traits BIGA-BIGO - values artistic, aesthetic experiences
W9BIGAO0K	BIG5 Personality traits BIGA-BIGO - is considerate and kind to almost everyone
W9BIGAO0L	BIG5 Personality traits BIGA-BIGO - does things efficiently
W9BIGAO0M	BIG5 Personality traits BIGA-BIGO - is reserved
W9BIGAO0N	BIG5 Personality traits BIGA-BIGO - is relaxed, handles stress well
W9BIGAO0O	BIG5 Personality traits BIGA-BIGO - has an active imagination
W9DOPEN	DV: OCEAN - Openness Subscale
W9DCONS	DV: OCEAN – Conscientiousness Subscale
W9DEXTRAC	DV: OCEAN – Extraversion Subscale

W9DAGREE	DV: OCEAN – Agreeableness Subscale
W9DNEUROT	DV: OCEAN – Neuroticism Subscale

Please see the Derived Variables User Guide for the details on sub-scales (openness, conscientiousness, agreeableness, neuroticism and extroversion) based on the 15 items above.

5.4.9 AUDIT-C

Bush K, Kivlahan DR, McDonell MB, et al (1998). The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. Ambulatory Care Quality Improvement Project (ACQUIP). Arch Intern Med. 158:1789-95.

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 (W9AUDIT). Scores of 5 or more are considered AUDIT-C positive and associated with increasing or higher risk drinking.

Variable Name	**Variable label**
W9AUDIT1	How often has a drink containing alcohol
W9AUDIT2	How many drinks containing alcohol has on a typical day of drinking
W9AUDIT6	How often had six or more drinks on one occasion in the past year
W9DAUDIT	DV: Alcohol Use Disorders Identification Test Consumption (AUDIT-C) scale

Response options are different for each of the three questions above. For W9AUDIT1 these are “Never, Monthly or less, 2-4 times a month, 2-3 times a week, 4 or more times a week”, for W9AUDIT2 these are “1-2 drinks, 3-4 drinks, 5-6 drinks,

7-9 drinks, 10 or more drinks”, and for W9AUDIT6 “Never, Less than monthly, Monthly, Weekly, Daily, or almost daily”.

The AUDIT-C items were also asked at Sweep 8.

6. Research Data

6.1 Licensing and data access

All datasets are available from the UK Data Service (UKDS).

All users of the data need to be registered with the UKDS. Details of how to do this are available at https://www.ukdataservice.ac.uk/get-data/how-to-access/registration.

Safeguarded data (EUL)

The dataset listed in Table 7 are available from the UKDS as safeguarded data, which can be downloaded from the UKDS once the End User Licence (EUL) access conditions have been accepted by the user.

Please refer to section 6.12 for information on how these data have been de-identified for safe sharing.

These safeguarded data exclude detailed data that are sensitive and/or present a high risk for disclosivity. This applies to: open-text responses, personal identifiers, variables which are considered potentially disclosive and variables considered sensitive in nature.

Controlled data (Secure Access)

These disclosive and/or sensitive Next Steps data can be accessed as controlled data from the UKDS SecureLab. Applicants wishing to access this data need to abide by the terms and conditions of the UKDS Secure Access licence.

Before gaining access, researchers must make an application detailing the intended analysis and provide a justification as to why this data is requested. Application guidance can be found at https://ukdataservice.ac.uk/find-data/access-conditions/secure-application-requirements/apply-to-access-non-ons-data/

6.2 Datasets and data structure

The list of the Next Steps Age 32 datasets available from the UKDS is shown below. Where applicable, the naming and structure of the datasets are consistent with the Age 25 Sweep.

Most of the questionnaire data is contained in NS9_2022_Main_Interview.This file is deposited in a flat (wide format) file. i.e., where one record exists for each cohort member (CM).

In addition, there are several hierarchical (long format) datasets, which contain multiple records for each cohort member. 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.

Table 7. List of safeguarded datasets (End User Licence)

Dataset name	Contents	Structure	Identifier(s)
NS9_2022_Main_Interview	Modules 1 to 8, 12 and cognitive assessment scores	Flat	NSID
NS9_2022_Partnerships	Relationship histories	Hierarchical	NSID, W9RELID
NS9_2022_Person_Grid	Details of members living in the same household as CM including children and partners	Hierarchical	NSID, GRIDID
NS9_2022_Children_With_NonResident_Parents	Details of non-resident parents of children	Hierarchical	NSID, GRIDID
NS9_2022_NonResident_Children	Details of non-resident children	Hierarchical	NSID, GRIDID
NS9_2022_Pregnancy_History	Pregnancy histories	Hierarchical	NSID, W9PREGID W9CHILDNO

NS9_2022_Benefits	Details of individual benefits including unfolding brackets	Hierarchical	NSID, W9BENID
NS9_2022_Activity_History	Activities and Employment histories	Hierarchical	NSID, W9HISTID
NS9_2022_Derived_variables	Derived variables	Flat	NSID
NS9_2022_Cognitive_Tests	Cognitive responses for practice and test trials	Hierarchical	NSID W9COGID
Next_Steps_Longitudinal_File	Sample, weighting and outcome variables for sweeps 1 to 9	Flat	NSID

Table 8. List of controlled datasets (Secure Access)

Name	Contents	Structure	Identifier(s)
NS9_2022_Main_Interview_Sensitive	National identity, long standing illness, SOC/SIC codes, university attended, dates	Flat	NSID
NS9_2022_Pregnancy_History_Sensitive	Type of fertility treatment received	Hierarchical	NSID, W9PREGID, W9CHILDNO, GRIDID

6.3 Data documentation

In addition to this User Guide, the following documentation accompanies the data deposited at the UKDS:

Table 9. Data documents

Name of the document	Content summary
Next Steps - Sweep 9 (Age 32) Questionnaire	This document provides the questions asked in the Age 32 sweep (the full/long version of the questionnaire), including details on any routing, and mode specific adjustments. It also reflects any changes that were made to the questionnaire after the Soft Launch.
Next Steps - Sweep 9 (Age 32) Online Short Mop-up Questionnaire	This document provides the short version of the questionnaire sent to non-respondents after the face-to-face phase (during the mop-up phase).
Next Steps - Sweep 9 Survey Technical Report	This document, authored by the implementing survey agency Ipsos, provides all technical details regarding the design and implementation of the survey.
Next Steps Age 32 Sweep (Sweep 9) Derived Variables Guide	This document provides information on new variables that were created by CLS following data collection, based on existing ones.

6.4 Identifiers

Individual identifiers

All datasets are primarily identified with the same research identifier (NSID) used for all Next Steps cohort data available at the UKDS.

Other identifiers

For hierarchical datasets, one or more secondary indices are given which uniquely identify a case. For example, in the person grid, the household member is uniquely identified by NSID GRIDID. In the pregnancy history grid each case is uniquely identified by NSID W9PREGID W9CHILDNO where W9PREGID denotes 1st, 2nd,.. pregnancy and W9CHILDNO denotes the details of each child within that pregnancy. The total number of pregnancies is denoted by W9PREGMANY. GRIDID facilitates matching to other person grids.

6.5 Variable description

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.

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 ‘W9’ 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**
W9CQUC0A – W9ACQUCB0V	Academic qualifications currently studying for
W9BIGAO0A – W9BIGAO0O	BIG5 Personality traits BIGA-BIGO

Derived variables in the dataset ‘NS9_2022_Derived_variables’ are given the prefix “W9D”.

Variable labels

Variable labels are based on the wording that can be found in the questionnaire documentation. Where necessary, labels have been modified to ensure they are comprehensible and accurate.

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”.

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.

6.6 Derived variables

Several 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.

Table 10. List of Derived variables

Geography		
	Variable Name	**Variable Label**
	W9DRGN	DV: Interview government office region
	W9DIMDD	DV: 2019 Index of Multiple Deprivation Decile
Household Relationships		
	Variable Name	**Variable Label**
	W9DAGEINT	DV: Age in months at interview
	W9DHSIZE	DV: Number of people currently living in household (inc CM)

	W9DCHNO	DV: Number of children currently living in household
	W9DCHOWNNO	DV: Number of own children currently living in household
	W9DCHPARNO	DV: Number of children of CM's current or previous partner in household
	W9DCHNO4	DV: Number of own children between 0 and 4
	W9DCHNO11	DV: Number of own children between 5 and 11
	W9DFATHER	DV: Whether CM's father in household
	W9DMOTHER	DV: Whether CM's mother in household
	W9DMARSTAT	DV: Legal marital status
	W9DCOHAB	DV: Whether has a cohabiting partner
	W9DPARTP	DV: Whether has a spouse or partner
Housing		
	Variable Name	**Variable Label**
	W9DTIMAD	DV: Time at current address (months)
	W9DTENURE	DV: Housing tenure
	W9DRENTFROM	DV: Who rents from
	W9DWHOTEN	DV: Whose name accommodation held in
Employment		
	Variable Name	**Variable Label**
	W9DACTIVITYC	DV: Current activity of CM - back coded
	W9DWRK	DV: Whether CM currently employed
	W9DEMPSZ	DV: Employment status/size of organisation for cohort member
	W9DWRKP	DV: Whether CM's partner currently employed
	W9DDACTIVITYP	DV: Current activity of CM's partner (derived)
	W9DWRKCP	DV: Combined labour market status
Finance		
	Variable Name	**Variable Label**
	W9DINCB	DV: Banded weekly income
	W9DBENE	DV: Whether cohort member or partner receives any benefits
	W9DBENE2	DV: Whether cohort member or partner receives any benefits (incl extra split)
Education		

	Variable Name	Variable Label
	W9DANVQH	DV: Highest NVQ level from an academic qualification reported in Age 32
	W9DHANVQH	DV: Highest NVQ level from a vocational qualification reported in Age 32
	W9DDEGP	DV: Whether achieved first degree or higher
	W9DRUSSELL	DV: Whether degree awarded by Russell Group University
Health		
	Variable Name	**Variable Label**
	W9DBMI	DV: Body mass index
	W9DBMICA	DV: Body mass index category
	W9DDISEA	DV: Disability classification Equality act (2010)
	W9DDISEU	DV: Disability classification EU-SILC
	W9DHGTM	DV: Height in metres – self reported
	W9DWGHTK	DV: Weight in kilograms – self reported
	W9DSMOKE	DV: Smoking habits
Identity		
	Variable Name	**Variable Label**
	W9DETHN6	DV: Ethnic group - 6 category census class
	W9DETHN8	DV: Ethnic group - 8 category census class
	W9DETHN11	DV: Ethnic group - 11 category census class
	W9DETHN15	DV: Ethnic group - Detailed
	W9DETHNP6	DV: Ethnic group of CM's partner - 6 category census class
	W9DETHNP8	DV: Ethnic group of CM's partner - 8 category census class
	W9DETHNP11	DV: Ethnic group of CM's partner - 11 category census class
Self-completion		
	Variable Name	**Variable Label**
	W9DAUDIT	DV: Alcohol Use Disorders Identification Test Consumption (AUDIT-C) scale
	W9DCANEVER	DV: Whether CM has ever tried cannabis
	W9DSEXEVER	DV: Whether CM has ever had sex
	W9DGHQSC	DV: General Health Questionnaire (GHQ12) score (Goldberg & Williams,1988

	W9DGAD2	Generalised Anxiety Disorder 2-item
	W9DPHQ2	Patient Health Questionnaire 2-item
	W9DGRIT18	Short GRIT scale
	W9DLONELINESS	UCLA loneliness 3 item
	W9DOPEN	DV: OCEAN – Openness Sub Scale
	W9DCONS	DV: OCEAN – Conscientiousness Sub Scale
	W9DEXTRAV	DV: OCEAN – Extraversion Sub Scale
	W9DAGREE	DV: OCEAN – Agreeableness Sub Scale
	W9DNEUROT	DV: OCEAN – Neuroticism Sub Scale
	W9DFINLIT3	DV: Financial Literacy – All three questions correct’
	W9DFINLITA	DV: Financial Literacy – Number of questions attempted’
	W9DFINLITC	DV: Financial Literacy – Number of correct answers’
	W9DFINLITR	DV: Financial Literacy- At least one question refused

6.7 Income and payment unfolding brackets

A feature of income or payment questions is the use of unfolding brackets for those cases where a respondent refuses or is unable to provide an exact answer. The unfolding brackets questions are designed to elicit a minimum and maximum value that define a range or “closed band” within which the actual value lies.

On entering the unfolding brackets, respondents are asked to say whether they have more, less or about the same as a particular value. This question is repeated using different values (which will be a lower or higher value depending on the answer to the preceding question). The procedure stops at the point when either: an upper and lower bound is provided; the respondent refuses or says “don’t know”; or the respondent places themselves in the top or bottom bracket.

The unfolding bracket questions are randomly ordered for each respondent. This will average any possible 'anchoring' effects (i.e. where people use the suggested figure as a reference point and adjust it to reach their answer) from the procedure across

the distribution. The bracket values are selected based on the density of the underlying financial variable.

6.8 Person grids

The person grid is comprised of five separate loops within the CAPI questionnaire; partner grid, two child grids (children reported at Age 25 Sweep and additional children reported in this sweep), and two ‘other’ household members grids (household members that are not partners or children reported at Age 25 Sweep and anyone not reported in the four other grids). Together these cover all possible household members at the time of interview as well as previous household members who have since left. The information is supplemented with feed-forward information from prior sweeps where questions were unasked.

6.9 Missing values

Missing values are consistently labelled as follows:

-9 = Refusal

-8 = ‘Don't Know’ (survey) or ‘Not codeable’ for derived variables

-1 = Item not applicable

-2 = Script Error

-3 = Not asked at fieldwork stage

In the Age 32 Sweep –3 has been reserved for questions not asked in the short mop-up questionnaire and –2 has been used to flag questions unanswered as a result of the CAPI script error described in 6.11.

The value –1 is also used for missing responses to questions which study members would not have been asked if they only partially completed the survey. For derived variables -8 is typically reserved for ‘Not codeable’ values, where there is insufficient data for the variable to be derived.

Additionally, the DATA_AVAILABILITY variable in the Next_Steps_Longitudinal_File describes cohort members for whom no data is available from any sweep of the study, following requests to delete their survey data.

6.10 Data cleaning

Manual edits were carried out on the cases affected by the CAPI script below. Individuals were verified using data collected in this sweep and compared to the feed-forward data. Those who could be verified were moved to the correct loop. Edits have been made to W9MORECH, W9MCHMANY, W9MORE and W9MHMANY to flag these cases. In addition, cleaning of identifiers has taken place for existing children identified in other datasets.

In the partnerships file where NRANY = 2 & NRLIVEBM <> -1 these cases have been set to –1.

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 flag variables have been added and included in the derived variables dataset.

6.11 Data errors and inconsistencies

Users should be aware of the following data corrections and details:

Following an error in the CAPI script, hierarchical feedforward data was not pulled through and verified as intended. This affected the child loops and other household loops directly. There was also an impact on the pregnancy loop, non-resident parent and non-resident children loops. In the child loop 189 cases had feedforward data but were routed to the new child loop. In the household loop 183 cases had feedforward data but were routed to the new household loop. A flag is included in the person grid to denote which household members were affected.

Additionally, during the Soft Launch, due to a script related error, part of the data collected about occupation coding were not recorded for some of the online interviews. This meant that occupation codes could only be produced by office coding of variables that were not affected – a subset of variables needed for office coding. The variable W9SOCERROR identifies these cases. In the Next Steps Sweep 9 Survey Technical Report Section 5.4.2 and Appendix 10.1 provides details of the issue and variables that weren't recorded, and Section 9.2.3 provides an evaluation of the coding done for these cases.

6.12 Data de-identification

In addition to the pseudo-anonymisation, all text variables that contained detailed information provided by the respondents have been removed from the research dataset. This includes job titles, job descriptions, exact names of education institutions, town name, postcodes and the final open-ended question.

In this deposit the original SOC and SIC codes have been merged and truncated to three digits in line with previous sweeps. The variable denoting number of rooms in the house has been top-coded to a maximum of 12, the JACS code for degree obtained has been truncated to 2 digits and the long standing conditions and illnesses questions are represented at the highest category. The original version of the variables are either available under Secure Access or can be requested separately.

Table 11. Variables edited for de-identification purposes

Variable name	Variable label
W9DNUMROOMS	Number of rooms in current home (top-coded)
W9DSUBDEG	Subject of degree (JACS3 2 digit code)
W9CURRJOB2020	DV: Current Job SOC2020 (3 digits)
W9CURRJOB2010	DV: Current Job SOC2010 (3 digits)
W9AGE25JOB2020	DV: Age 25 Job SOC2020 (3 digits)
W9AGE25JOB2010	DV: Age 25 Job SOC2010 (3 digits)
W9FIRSTJOB2020	DV: First Job SOC2020 (3 digits)
W9FIRSTJOB2010	DV: First Job SOC2010 (3 digits)
W9PARTNERJOB2020	DV: Partner's Job SOC2020 (3 digits)
W9PARTNERJOB2010	DV: Partner's Job SOC2010 (3 digits)
W9DLOILCOND0A - W9DLOILCONDAC	Conditions expecting to last 12 months: Respiratory problems - Conditions expecting to last 12 months: Any other condition

6.13 Output Disclosure Control

Data included in Table 8 is only available via the UKDS Secure Lab. Access to this data is controlled and the UK Data Service will always perform a certain level of disclosure control on the outputs generated by researchers, as outlined in their SDC Handbook which can be downloaded from: https://securedatagroup.org/sdc-handbook/

The two UK Data Service Secure Lab rules of thumb that will be applied to all outputs are:

- Threshold rule: No cells should contain less than 10 observations
- Dominance rule: No observation should dominate the data to a huge extent

7. Response patterns and weights

7.1 Response patterns

The issued sample for first sweep of Next Steps 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 (Black Caribbean and Black African pupils, selected from the original [non-responding] school sample), taking the total number of cohort members who had taken part in the study up to 16,122.

7,279 cohort members participated at Sweep 9 (45.1% of the total sample). Sample sizes and response rates (as a proportion of all participants) for each sweep are shown in Figure 2.

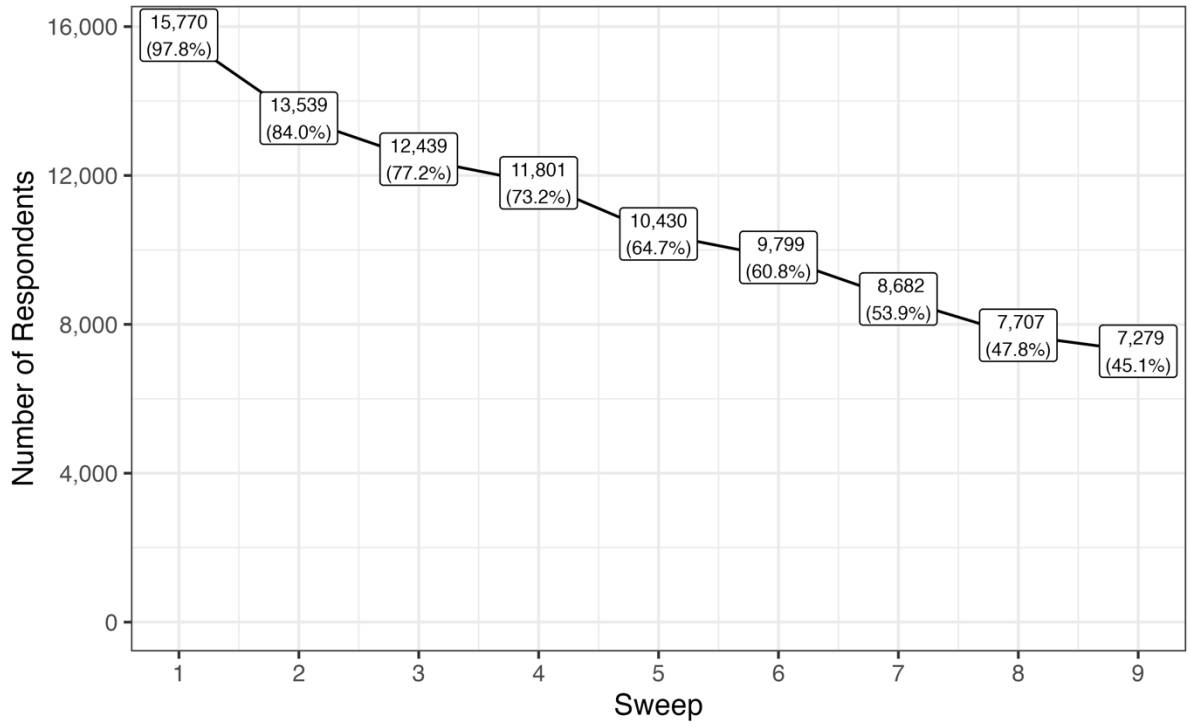


Figure 2. Sample size and response rates by sweep (productive participants)

Note: Response rates calculated as proportion of total (n = 16,122) sample, and thus includes the boost sample in the denominator for Sweeps 1-3, when these participants were not eligible for inclusion in the study.

In Sweeps 2-7 only those who had participated in the prior sweep were invited to participate[4]. This was not the case in Sweeps 8 and 9 where all participants who had ever previously participated were invited to take part unless they had died, permanently withdrawn or become permanently untraced. As such, participants at Sweeps 8 and 9 included those who had not taken part for some time.

In Figure 3, Panel A shows the proportion of participants at a given sweep who had participated at all prior sweeps.

Panel B shows the proportion of Sweep 9 participants who participated at a given prior sweep (useful for understanding missingness in data from prior sweeps).

Panel C shows the proportion of participants at a given sweep who participated at Sweep 9 (useful for understanding attrition).

[4] There were two exceptions to this: In Sweep 5 a small number Sweep 4 non-respondents who asked to be re-included were also invited to take part. In Sweep 6 a substantial proportion of non-respondents who were issued at Sweep 5 were re-invited to take part.

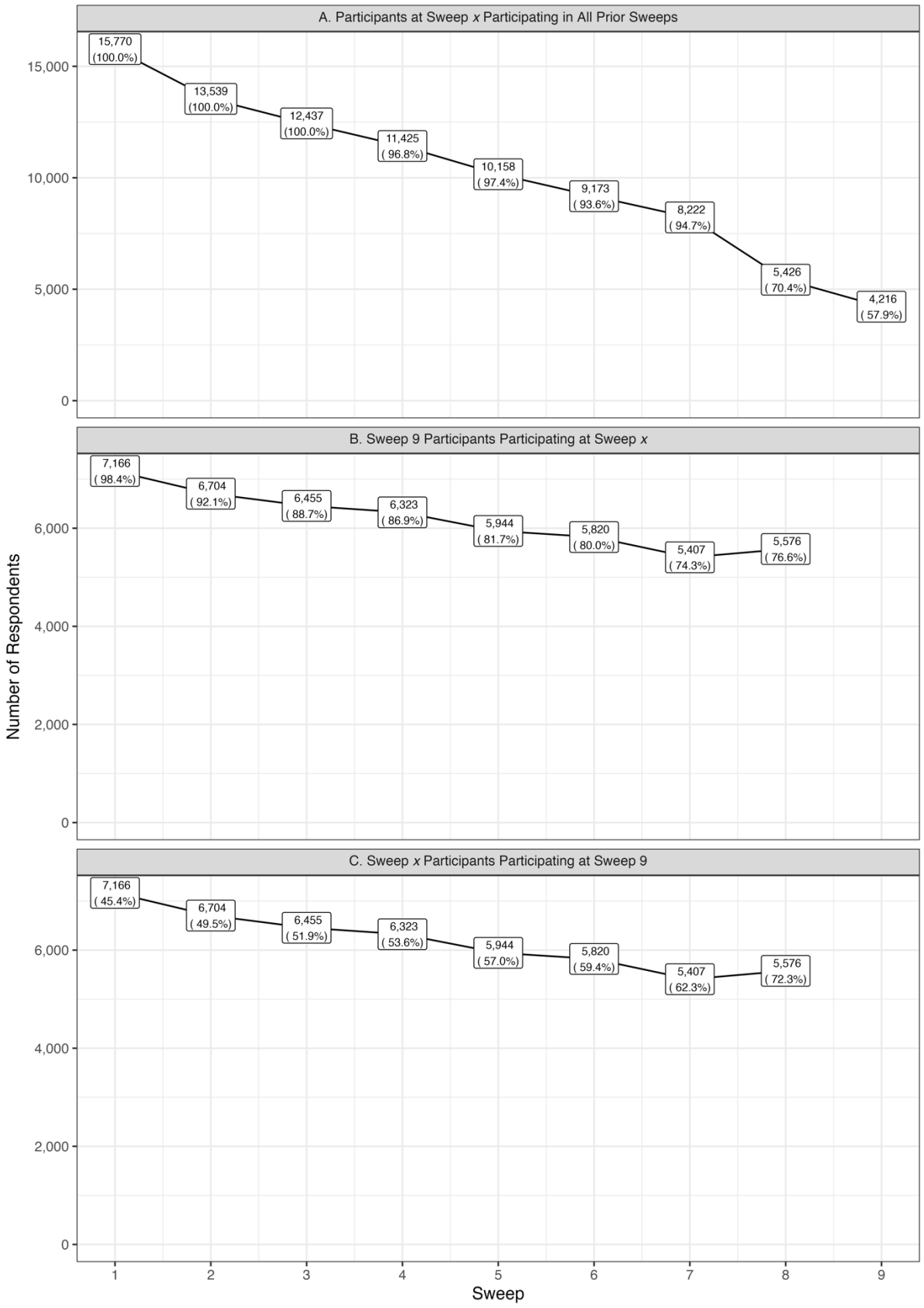


Figure 3. Response rates by sweep

Note: Although the ethnic minority boost sample was added in Sweep 4, respondents to this sweep has been included in the denominator in Panel B, which is why the proportion of Sweep 9 respondents who took part in Sweep 1 is <100%.

7.2 Predicting response in Sweep 9 and weights

7.2.1 Introduction

In longitudinal studies like Next Steps, it is inevitable that some users will drop-out of the study. Attrition can mean less statistical power but can also introduce bias as respondents often differ systematically from non-respondents. There are several approaches that can be used to handle missing data, including multiple imputation (MI), weighting, and full-information maximum likelihood (FIML). Elsewhere, CLS has developed comprehensive advice on dealing with missing data in its cohort studies (Silverwood et al. 2024). Here we discuss, in abbreviated form, the derivation and implementation of non-response weights that are provided with the Next Steps Age 32 Sweep data. More detailed information is provided in the Appendix.

7.2.2 Target populations and response definitions

Four separate non-response weights are provided with the data (W9FINWTALLA, W9FINWTLONGA, W9FINWTALLB, W9FINWTLONGB), reflecting different target populations (the population the weights attempt to recover) and different definitions of 'response'.

Target Populations:

1. **Target Population A:** People who were in Year 9 in England in February 2004 and were living in England and not in prison or on probation at Age 32 in 2022/23 (n = 15,527).
2. **Target Population B:** People who were in Year 9 in England in February 2004, and were alive and not in prison or on probation at Age 32 in 2022/23 (n = 16,012)

These target populations reflect differences in eligibility in terms of residence for the Age 32 Sweep (see Sections 3 and 4). Note, in practice there are only small

differences in the size of the target populations, and results may be similar regardless of which target population is used when selecting a weight.

Definitions of Response:

1. All Age 32 Participants (n = 7,279)
2. Age 32 Participants of the full/long version of the questionnaire (n = 6,791)

These definitions reflect differences in the content of the full questionnaire and the short version used (for some participants) in the mop-up survey. Participants who completed the short mop-up survey have missing values for questions which were only included in the full questionnaire. Accordingly, analysts using variables that were only collected in the full questionnaire should consider using the relevant weight to reduce non-response bias.

The variables containing the calculated weights are as follows:

Variable name	**Variable label**
W9FINWTALLA	W9 Final Weight: Full/long surveys AND short mop-up surveys (Resident in England)
W9FINWTLONGA	W9 Final Weight: Full/long survey only (Resident in England
W9FINWTALLB	W9 Final Weight: Full/long surveys AND short mop-up surveys (Resident in England and elsewhere)
W9FINWTLONGB	W9 Final Weight: Full/long survey only (Resident in England and elsewhere)

7.2.3 Derivation of Non-Response Weights

The non-response weights were created using the following procedure, repeated for each combination of target population and definition of response. First, as not all Sweep 9 respondents participated at Sweep 8, we split the sample according to prior sweeps responded at to reduce missingness in variables used to predict response: for those who participated at Sweep 8, we could use information from this most recent available information, but for those who did not, we had to rely only on information collected earlier. This included splitting off the boost sample (n = 352), who only entered the study at Sweep 4.

Second, in each of these samples (except the boost sample) we created random forest models to obtain predicted probabilities of response using information collected at earlier sweeps. We took a relatively agnostic approach to selecting predictors and thus included a large set of potential predictors in models. Random forest models – a machine learning approach – were used to predict response as these can handle large numbers of (collinear) predictor variables, subsetting variables automatically based on their ability to predict response. For the boost sample, given the smaller sample size, we used linear regression with factor analysis first used to create (five) predictors capturing a sizeable proportion of the variation in a large number of Sweep 4 predictors (sex was included as a separate predictor of response, given the importance of this variable for many analyses).

Third, we converted the predicted probability to weights by taking their inverse – higher values reflect a lower probability of response and thus a responding individual with a higher weight is used to reflect a larger section of the target population.

Fourth, as Next Steps used a stratified sampling design and there was non-response to the initial survey as well as subsequent attrition, we created 'final' weights by combining our Sweep 9 response weight with final weights from prior sweeps that have been previously created for the survey. The prior final weights that were used in this step depended on the sweep the sample used to create the response weight was last (uniformly) observed at (e.g., Sweep 4 for the boost sample).

Final weights for previous sweeps were created to make the sample representative of the initial target population (individuals who attended Year 9 of secondary school in England in 2003/04), a definition that diverges slightly from the definitions of the target populations used here. However, combining Sweep 9 response weights with previous sweeps' final weights was necessary due to lack of other data on the target populations. Nevertheless, as participants in Target Populations A or B comprise 95% or more of the full Next Steps sample, this is unlikely to introduce much bias in practice.

Sweep 9 final weights are supplied with the dataset as the variables W9FINWTALLA, W9FINWTLONGA, W9FINWTALLB, and W9FINWTLONGB. These have been scaled so that the sum of each set of weights equals the final achieved sample size for those in the respective target population. Note, we did not truncate

the weights (pre- or post-scaling or following combination with prior cross-sectional weights) as truncation appeared to lead to worse performance.

7.2.5 Implementation of Non-Response Weights

Data users should select a weight based on (a) the population they want their analysis to be representative of and (b) the variables used in their analysis. For instance, for policy questions using mental health data (which only appear in the full questionnaire), data users may consider participants resident in England in 2022/23 (Target Population A) as the relevant target population and full questionnaire response as the relevant definition of response. However, data users should consider whether creating their own bespoke weights is required. This may be the case if, for instance, the analysis to be undertaken only uses data from those who displayed a particular response pattern (e.g., participated in Sweep 1 and Sweep 9), or are otherwise a subset of all Sweep 9 respondents (e.g., those in employment, those with children).

Next Steps used a complex sampling design to recruit cohort members: cohort members were recruited from stratified samples of schools (approximately 30 students per school; see the LSYPE User Guide to Datasets: Wave 1 to Wave 7 for more detail). Analyses of Next Steps data should account for this complex sampling design by specifying primary sampling unit (PSU; SAMPPSU) and stratum (SAMPSTRATUM). In Stata, this can be achieved by first using `svyset` to specify the survey design, and then conducting analyses using the `svy` prefix, e.g.:

```
svyset SAMPPSU [pweight=W9FINWTALLA], strata(SAMPSTRATUM)
svy: proportion W9DSEX if W9DSEX >= 0, citype(agresti)
```

In R, the `survey` package can be used to specify complex survey design, e.g.:

```
library(survey)
ns_svy <- svydesign(id = ~ SAMPPSU, strata = ~ SAMPSTRATUM, weights
= ~ W9FINWTALLA, data = ns_w9)
svytable(~ W9DSEX, ns_svy)
```

See relevant Stata help files (StataCorp 2023) and `survey` package documentation (Lumley 2011) for more information on using survey data in Stata and R. (Users with experience of the `tidyverse` may also want to use the `srvyr` package, which provides similar functionality to `survey` but within a `tidy` framework.) Given attrition, in some cases, there may only be one PSU in a stratum, in which case users may consider allocating these observations to the modal stratum. Also, for some commands, users may find that certain functionalities have not been adapted for use with complex survey data. In this case, users may consider using survey weights without declaring complex survey design and noting this in the write up of their analyses. Users can also check whether accounting for complex survey design makes much difference in their particular analysis, by running analyses declaring and not declaring the complex design; as participants have now long left secondary school, observations at Age 32 may be independent (or at least less dependent) at the PSU level.

8. Mode effects

Sweep 9 of Next Steps used a sequential mixed mode design to lower costs and increase participation rates. An issue with mixed mode designs is the potential for responses to differ systematically between survey modes. For instance, the presentation of a survey item either orally or visually can influence responses, and sensitive information may be reported more accurately when given anonymously (e.g., by web survey compared with face-to-face interview). Differences in responses arising from differences in measurement between surveys modes are termed ‘mode effects’.

Unaccounted for, mode effects can generate bias in analyses, both for descriptive and inferential statistics. For instance, estimates of the change in mental health scores may reflect differences in the survey modes used.

Simply adding an indicator variable for survey mode into analyses of Next Steps Sweep 9 data may not be sufficient to remove bias as selection into mode was not random; participants who did not respond to initial invitations will are less likely to have completed the survey via web, and likely differ on a number of dimensions from those who responded at first contact. Observed differences between modes are a combination of mode effects and selection effects. Adding an indicator variable for mode does not account for this.

CLS will soon release documentation on handling mode effects in its cohort studies, including worked examples in R and Stata and a set of recommendations that we suggest researchers follow in their own analyses of CLS data.

9. References

Brown, M., Fitzsimons, E., Goodman, A., Peters, A., Ploubidis, G.B., Sanchez, A., Silverwood, R., Smith, K. (2021) *COVID-19 Survey in Five National Longitudinal Studies: Waves 1, 2 and 3 User Guide (Version 4)*. London: UCL Centre for Longitudinal Studies and MRC Unit for Lifelong Health and Ageing. https://cls.ucl.ac.uk/wp-content/uploads/2017/02/UCL-Cohorts-COVID-19-Survey-user-guide.pdf

Calderwood, L., Peycheva, D., Henderson, M., Silverwood, R., Mostafa, T., Rihal, S. (2021) *Next Steps: Sweep 8 – Age 25 User Guide (3rd edition)*. London: UCL Centre for Longitudinal Studies. https://cls.ucl.ac.uk/wp-content/uploads/2020/09/NextSteps_Age_25_Survey_User_Guide.pdf

Centre for Longitudinal Studies. (2024). Next Steps – Sweep 9 (Age 32) Online Short Mop-up Questionnaire. https://cls.ucl.ac.uk/wp-content/uploads/2017/02/Next-Steps-Age-32-Online-Short-Mop-Up-Questionnaire.pdf

Centre for Longitudinal Studies. (2024). *Next Steps – Sweep 9 (Age 32) Questionnaire*. https://cls.ucl.ac.uk/wp-content/uploads/2023/07/Next-Steps-Age-32-Survey-Questionnaire-2.pdf

Department for Education. (2011). *LSYPE User Guide to the Datasets: Wave 1 to Wave 7*. https://doc.ukdataservice.ac.uk/doc/5545/mrdoc/pdf/lsype_user_guide_wave_1_to_wave_7.pdf

Ipsos. (2024) *Next Steps – Sweep 9 Survey Technical Report*. London, Ipsos.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in R*. Springer.

Lumley, T. (2011). *Complex surveys: a guide to analysis using R*. John Wiley & Sons.

Silverwood, R., Narayanan, M., Dodgeon, B., Katsoulis, M., Ploubidis, G. (2024) *Handling missing data in the CLS cohort studies: User guide*. London: UCL Centre for Longitudinal Studies.

Singh, S., Strong, R. W., Jung, L., Li, F. H., Grinspoon, L., Scheuer, L. S., Passell, E. J., Martini, P., Chaytor, N., Soble, J. R., & Germine, L. (2021). The TestMyBrain Digital Neuropsychology Toolkit: Development and Psychometric Characteristics. *Journal of clinical and experimental neuropsychology*, *43*(8), 786–795. https://doi.org/10.1080/13803395.2021.2002269

StataCorp. 2023. *Stata 18 Survey Data Reference Manual*. College Station, TX: Stata Press.

Appendix: Derivation of Sweep 9 final weights

Introduction

In this Appendix, we provide detail on the derivation of the four survey weights. Abbreviated information is provided in Section 7.

As is common in longitudinal studies, there has been attrition from Next Steps over time; the total participating sample at Sweep 9 (n = 7,279) represents 45.1% of the total recruited sample. Non-response can cause bias in analyses as participants continuing in a study may differ along a number of dimensions compared with those who drop-out. For instance, males were less likely to participate at Sweep 9 of Next Steps than females.

One strategy for accounting for non-response bias is weighting. Respondents with underrepresented characteristics are given higher weight in analyses so that the (weighted) population reflects the population from which respondents are drawn, at least with respect to measured characteristics.

The Sweep 9 data contain four survey weights designed to account for differential response at Sweep 9. Each of these weights reflects (a) a different 'target' population (the population the weight is designed to recover), given that only some individuals were eligible to participate in the face-to-face survey, and (b) a different definition of Sweep 9 'response', given that some participants at Sweep 9 only answered a short mop-up survey questionnaire, which did not contain many of the items collected in the full/long version. Specifically:

Target Populations:

1. **Target Population A:** People who were in Year 9 in England in February 2004 and were living in England and not in prison or on probation at Age 32 in 2022/23 (n = 15,527).
2. **Target Population B:** People who were in Year 9 in England in February 2004, and were alive and not in prison or on probation at Age 32 in 2022/23 (n = 16,012)

Definitions of Response:

1. All Age 32 Participants (n = 7,279)
2. Age 32 Participants of the full/long version of the questionnaire (n = 6,791)

Derivation of survey weights

Creating each survey weight involved several steps, which we repeated for each combination of target population and definition of response to obtain the four weights in total.

1. Estimate a participant's probability of responding at Sweep 9 using available information obtained at Sweeps 1-8.
2. Take the inverse of this probability to create a Sweep 9 non-response weight among Sweep 9 participants.
3. Combine (by multiplication) this Sweep 9 non-response weight with 'final' weights from prior sweeps to obtain a Sweep 9 final weight that accounts for (a) prior non-response at Sweeps 1-8 and (b) non-random recruitment into Next Steps due to the complex survey design and initial non-response.

These steps are explained in greater detail in what follows.

Step 1: Estimating the probability of responding at Sweep 9

Next Steps contains very rich data on participants' lives, and we utilised this to estimate probabilities of response at Sweep 9. However, as there has been a non-monotonic response in Next Steps (see Section 7), information from each prior sweep was not available for all participants at Sweep 9; importantly, data from Sweeps 1-3 was not available for the boost sample who entered at Sweep 4, and data from Sweep 8 was not available for participants who had last participated at Sweep 7 or earlier (24.4% of Sweep 9 participants).

To make the most of the available data while reducing item missingness to acceptable levels, we split the sample into four groups, each defined by a participants appearance in a specific set of sweeps. This allowed us to use information from these sweeps as predictors of Sweep 9 response for that group.

The groups are displayed in Appendix Table 1 below. As noted above, weighted were created for each combination of target population and definition of Sweep 9 response, so this splitting of the sample into the below groups was carried out after first defining the target population and definition of Sweep 9 response. (Sample sizes given below are for Target Population A with response defined as answering any questionnaire at Sweep 9, full or short.)

Appendix Table 1. List of groups used to predict Sweep 9 response

Group	Eligible Sample	Relevant Respondents	Size	Predictor Sweeps
I	Original sample (entry at Sweep 1) who participated at Sweep 8	Eligible sample who responded at Sweep 9	Eligible Sample = 7,569 Relevant Respondents = 5,495 (72.6%)	1, 8
II	Original sample (entry at Sweep 1) who participated at Sweep 7	Eligible sample who responded at Sweep 9 but not Sweep 8	Eligible Sample = 8,494 Relevant Respondents = 984 (11.6%)	1, 2, 3, 6, 7
III	Original sample (entry at Sweep 1)	Eligible sample who responded at Sweep 9 but not Sweep 7 or Sweep 8	Eligible Sample = 15,770 Relevant Respondents = 687 (4.4%)	1
IV	Boost sample (entry at Sweep 4)	Eligible sample who responded at Sweep 9	Eligible Sample = 352 Relevant Respondents = 113 (32.1%)	4

Note: Sample sizes in the 'Size' column are for the Target B population with Sweep 9 response defined as any questionnaire (full or short).

Response models, from which the probability of response was estimated, were constructed for each of these four groups separately. The 'Eligible Sample' defined the set of participants included in the response prediction model (from a given target population), while 'Relevant Respondents' defined the set of positive cases (given a definition of Sweep 9 response – e.g., full questionnaire only).

The phrase 'Relevant Respondents' is important: a complexity is that the eligible samples in Groups I-III overlap so a given Sweep 9 respondent could appear in

multiple eligible samples (e.g., a participant at Sweeps 1, 7 and 8 would appear in Groups I, II, and III). 'Relevant Respondents' were defined so that each Sweep 9 respondent was classified as a 'Relevant Respondent' only once. This ensure that, for each Sweep 9 respondent, there was only one response model where they represented a positive case; thus, when we derived weights by taking the inverse of the response probability and recombined positive cases across groups (Step 2), the aggregate sample was all Sweep 9 respondents, with no Sweep 9 respondent appearing more than once.[5]

For each group, we selected a candidate set of predictors from sweeps in which a high proportion of the eligible sample were observed. For Group I, this was Sweeps 1 and 8; for Group II, this was Sweeps 1, 2, 3, 6, 7; for Group III, this was Sweep 1; and for Group IV, this was Sweep 4. The list of predictors from each Sweep is shown in Appendix Table 2.

Appendix Table 2. List of predictors used from previous Next Steps sweeps

Sweep	Text
1	Freq. alcohol use (Never; < Every few months; Every few months; Once a month; 2-3 times per month; 1+ per week); Freq. argue w/ main parent (Most days; More than once a week; Less than once a week; Hardly ever; Or never); Bullied in last 12 months (No; Yes); Ever tried cannabis (No; Yes); Computer in household (No; Yes); Disability or long term illness (None; Yes, schooling not affected; Yes, schooling affected); Ethnic group (White; Mixed; Indian; Pakistani; Bangladeshi; Black Caribbean; Black African; Other); Family type (Married couple; Cohabiting couple; Lone father; Lone mother; No parents in the household); Main household language (English; Other; Bilingual); Siblings in household (Range: 0-7); Does homework in term-time week (Yes; No); Attends independent school (No; Yes); Internet access at home (No; Yes); Educational intentions after

[5] Another way of thinking about this is that, as only one response probabilities could be used when create a final weight for a Sweep 9 respondent, if a Sweep 9 respondent was classified as a positive case in two or more groups, some of the response probabilities would need to be discarded. If some of the response probabilities were discarded, in aggregate, retained weights derived from these response probabilities would not reflect the eligible sample.

	Year 11 (Stay in Education; Leave Education); Mother's age (Range: 18-97); Family social class (NS-SEC: Higher Managerial and professional occupations; Lower managerial and professional occupations; Intermediate occupations; Small employers and own account workers; Lower supervisory and technical occupations; Semi-routine occupations; Routine occupations; Never worked/long term unemployed); Highest parental qualification (Degree; HE Below Degree; A Level; GCSE grades A-C; Level 1 or Below; None / Other); Police contact about CM behaviour (No Police Contact; Police Contact); Risk factor scale (Range: 0-8); Attitude to school scale (Range: 0-48); Special educational needs (No; Yes); Sex (Male; Female); Ever smoke cigarettes (No; Yes); Household tenure (Owned; Mortgage; Rent / Other); Played truant in last 12 months (No; Yes)
2	Region of residence (North East; North West; Yorkshire and The Humber; East Midlands; West Midlands; East of England; London; South East; South West); Risk factor scale (Range: 0-8); Attitude to school (Range: 0-48)
3	Risk factor scale (Range: 0-8); Attitude to school (Range: 0-48)
4	A-Levels being studied (No A-Levels; A-Levels); Ever drank alcohol (No; Yes); Birth country (United Kingdom; Elsewhere); Ever tried cannabis (No; Yes); Ethnicity (Black Caribbean; Black African; Other); Family type (Two Parent; One/No Parent); Number of GCSEs, Grade C or Above (Range: 0-10); GHQ-12 Caseness score (Range: 0-12); Self-rated health (Very Good; Not Very Good); Children in household (Range: 0-4); Interview month (Range: 6-10); Internet in household (No; Yes); Carried knife in last 12 months (No; Yes); Mother's age (Range: 22-76); Parental highest qualification (HE; Below HE; None); Contact with police about CM behaviour (No Police Contact; Police Contact); Attitude to school (Range: 0-20); Sex (Male; Female); Household tenure (Own; Rent / Other); Played truant (No Truancy; Truancy); Economic activity (Employed; Not Employed)
6	Economic activity (University; Other Education; Work; Training; Other); Ever tried cannabis (No; Yes); Disability or longstanding illness (No; Yes);

	Unpaid carer (No; Yes); Has child (No; Yes); Interview month (Range: 5-9); Survey mode (Web; Telephone; Face-to-Face); # close friends (None; 1; 2-3; 4-5; 6-9; 10 or more); Ever taken drugs (No; Yes); In relationship (No; Yes); Ever had sexual intercourse (Yes; No); Freq. doing sport (Most days; More than once a week; Once a week; Less than once a week; Hardly ever / Never)
7	Economic activity (Full Time Education; Non-Routine Work; Routine Work; Never Worked); Ever taken cannabis (No; Yes); Health problem or disability (No; Yes); Donates to charity (Yes; No); Has child (No; Yes); Life satisfaction (Range: 1-5); Survey mode (Web; Telephone; Face-to-Face); # close friends (None; 1; 2 - 3; 4 - 5; 6 - 9; 10 or more); Ever taken drugs (No; Yes); In relationship (No; Yes); Ever had sexual intercourse (Yes; Never); Freq. unpaid help (No; One Off; 1-2 times per year; Every couple of months; 1-2 times per month; 1-2 times per week); Voted at 2010 General Election (Yes, voted; No)
8	Economic activity (Work; Education / Training; Unemployment; Inactivity); AUDIT alcohol use (Range: 0-12); Body mass index (Range: 13.3-67.2); Highest qualification (NVQ Level 1; NVQ Level 2; NVQ Level 3; NVQ Level 4; NVQ Level 5; None/Other); GHQ-12 Caseness score (Range: 0-12); Longstanding illness (No; Yes); Interview Date (Range: 2015-08-01-2016-09-01); Freq. civic activity and volunteering (Range: 0-12); (Lives Alone; Lives with One Parent; Lives with Two Parents); (Single; In Relationship); (Web; Telephone; Face-to-Face); Number of children (No Children; 1; 2; 3+); Number of drugs tried (Range: 0-9); Patience rating (Range: 0-10); Interest in politics (Very interested; Fairly interested; Not very interested; Not at all interested); Freq. religious attendance (Once a week or more; Once a month or more; Sometimes but less than once a month; Never or very rarely); Risk tolerance (Range: 0-10); Self-rated health (Excellent; Very good; Good; Fair; Poor); Ever had sex (Yes; No); Hours sleep per night (Range: 4-10); Smoking status (Never; Ex-Smoker; Occasional; Daily); Social network use (None; 1 Hour; 2 Hours; 3 Hours; 4 Hours; 5 Hours; 6+ Hours); Trust rating (Range: 0-10)

For Groups I-III, we adopted a relatively inclusive approach to selecting predictors of response as we estimated response models using random forest modelling. The random forest algorithm by design selects variables based on predictive ability and can handle large numbers of correlated predictor variables and non-linear relationships between predictors and response (James et al. 2013). For Group IV, the sample size was much smaller (n_{max} = 352), so we instead used factor analysis for mixed data (FAMD) to extract five factors (explaining ~ 40% of the variance) from the candidate predictor variables (excluding sex) and then used logistic regression to predict response with these five factors plus sex added as predictors. This approach was chosen to adopt a relatively agnostic approach to predictor selection but also to explicitly balance upon sex, given the importance of this variable for many analyses.

The non-response models were estimated using multiply imputed data to address item-level missingness in the predictor variables (random forest algorithm, 32 imputed datasets).

Step 2: Creating a Sweep 9 Non-Response Weight from the Response Models

From each non-response models, we extracted predicted probabilities of response and took the inverse of these probabilities to create a response weight among relevant respondents. We then appended these data together to obtain a dataset will all Sweep 9 participants, with each participant appearing only once (see above). Weights were calculated in each of the 32 imputed datasets separately, then averaged.

Step 3: Combining with Prior Final Weights to Create Sweep 9 Final Weights

As Next Steps used a stratified sampling design and there was non-response to the initial survey as well as subsequent attrition, these Sweep 9 non-response weights alone were not sufficient to make the sample representative of the target populations. Accordingly, we created ‘final’ weights by combining our non-response weights with final weights from prior sweeps that have previously been created for the data. The prior sweep used depended on the group [I-IV] the participant belonged to. For Group I, this was Sweep 8; for Group II, Sweep 7; Group III, Sweep 1; and Group IV, Sweep 4.

Weights for previous sweeps were created to make the sample representative of the initial target population (individuals who attended Year 9 of secondary school in England in 2003/04), a definition that diverges slightly from the definitions of the target populations used here. Combining with previous sweeps' final weights was necessary due to lack of other data on the target populations. Nevertheless, as participants in Target Populations A or B comprise 95% or more of the full Next Steps sample, this is unlikely to introduce much bias in practice.

Sweep 9 final weights are supplied with the dataset as the variables W9FINWTALLA, W9FINWTLONGA, W9FINWTALLB, and W9FINWTLONGB. These have been scaled so that the resulting weights so that the sum of each set of weights equalled the final achieved sample size for those in the respective target population. Note, we did not truncate the weights (pre- or post-scaling or following combination with prior cross-sectional weights) as truncation appeared to lead to worse performance (see next section).

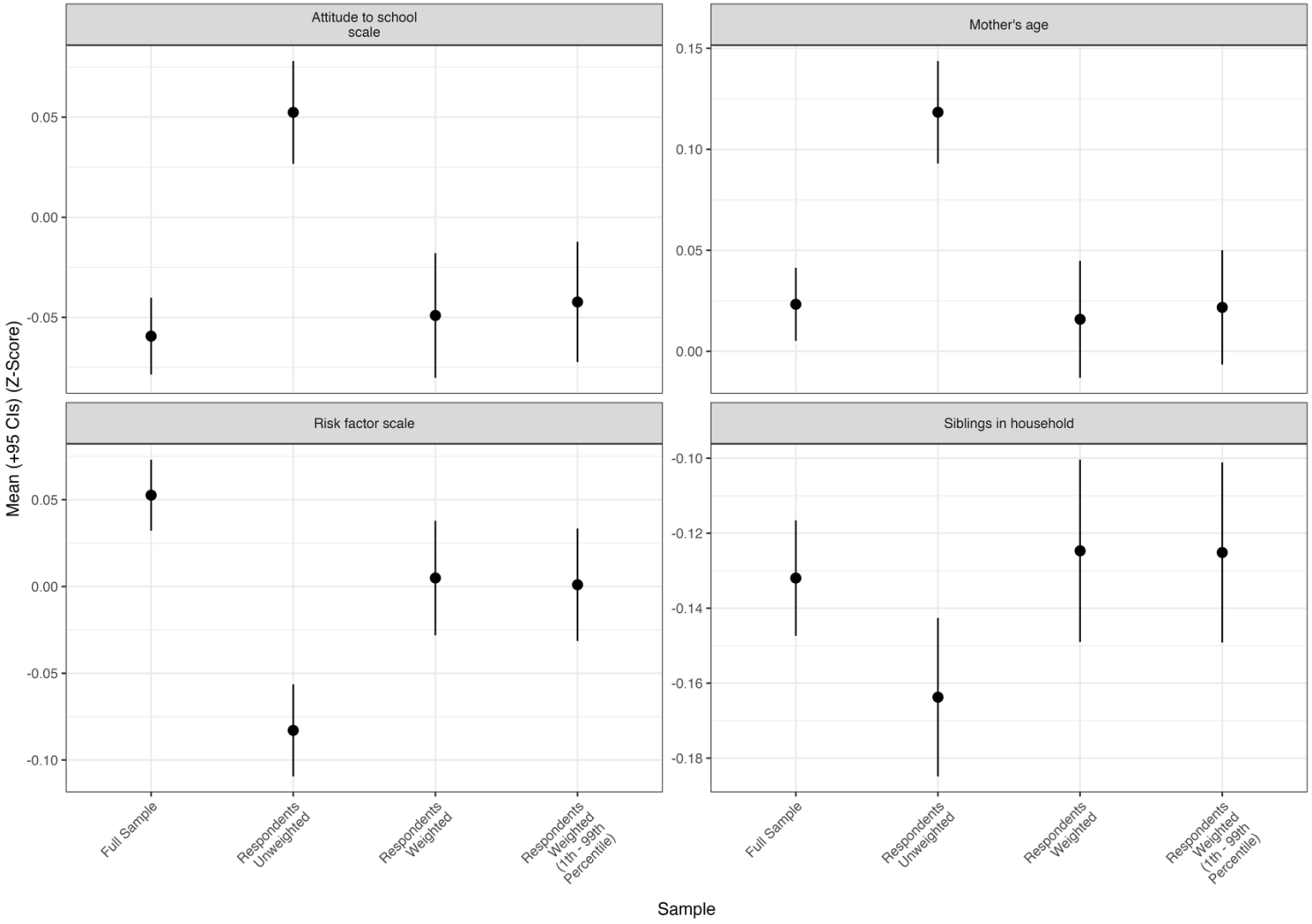
Effectiveness of weights

To examine the effectiveness of the derived Sweep 9 non-response weights in restoring sample representativeness we considered the distribution of the predictor variables (means for continuous variables, proportions for categorial variables) in Sweep 1 (original sample), Sweep 8 (original sample participating at this sweep) and Sweep 4 (boost sample). For brevity, plots for only a selection of results for W9FINWTALLA (Target Population A; any questionnaire at Sweep 9) are shown below; similar results were obtained for other three weights. The extent of bias is identified as differences in the mean (for continuous variables) or proportions (for categorical variables) among the (weighted) Sweep 9 participants compared with the unweighted sample from a prior sweep. Unweighted figures for the Sweep 9 participants are also provided for context, as well as figures when Sweep 9 weights were truncated at the 1st and 99th centiles of the distribution to reduce the influence of large weights.

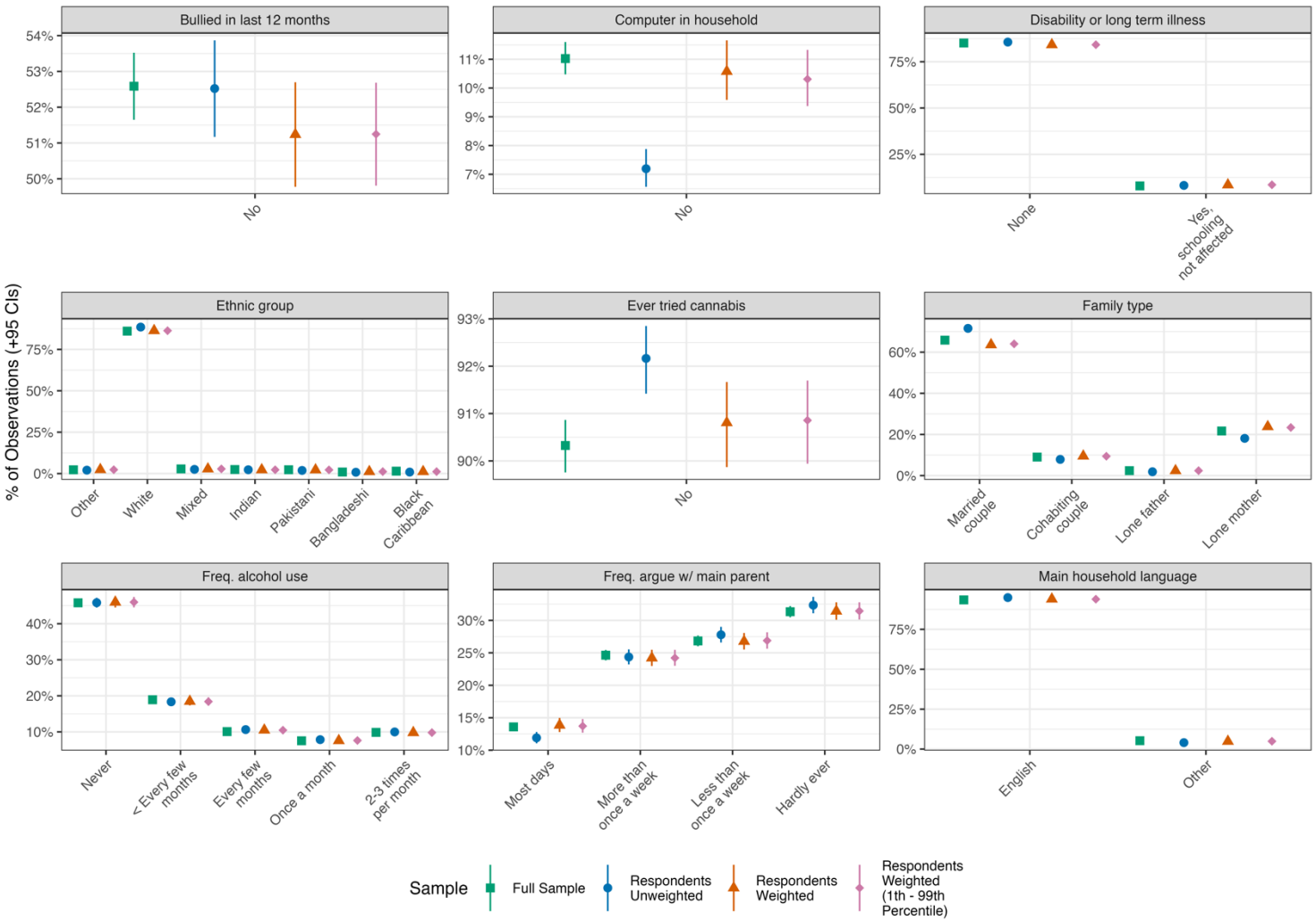
Weighting was effective for reducing bias and recovering initial distributions for some variables, but not others, though confidence intervals typically overlapped.

Specifically, for the variables measured in Sweep 1 for the original cohort members, weighting was successful in recovering initial sample averages means/proportions for, among other things, maternal age, family social class, household tenure, and educational intentions and special educational needs (Appendix Figures 1-4). Bias in the sex, playing truant, and ever smoking distributions was reduced but still present, which with bias increased for a small number of variables, including bullying victimisation. Performance for variables measured at Sweep 8 was typically worse than those measured at Sweep 1 (Appendix Figures 5-7).

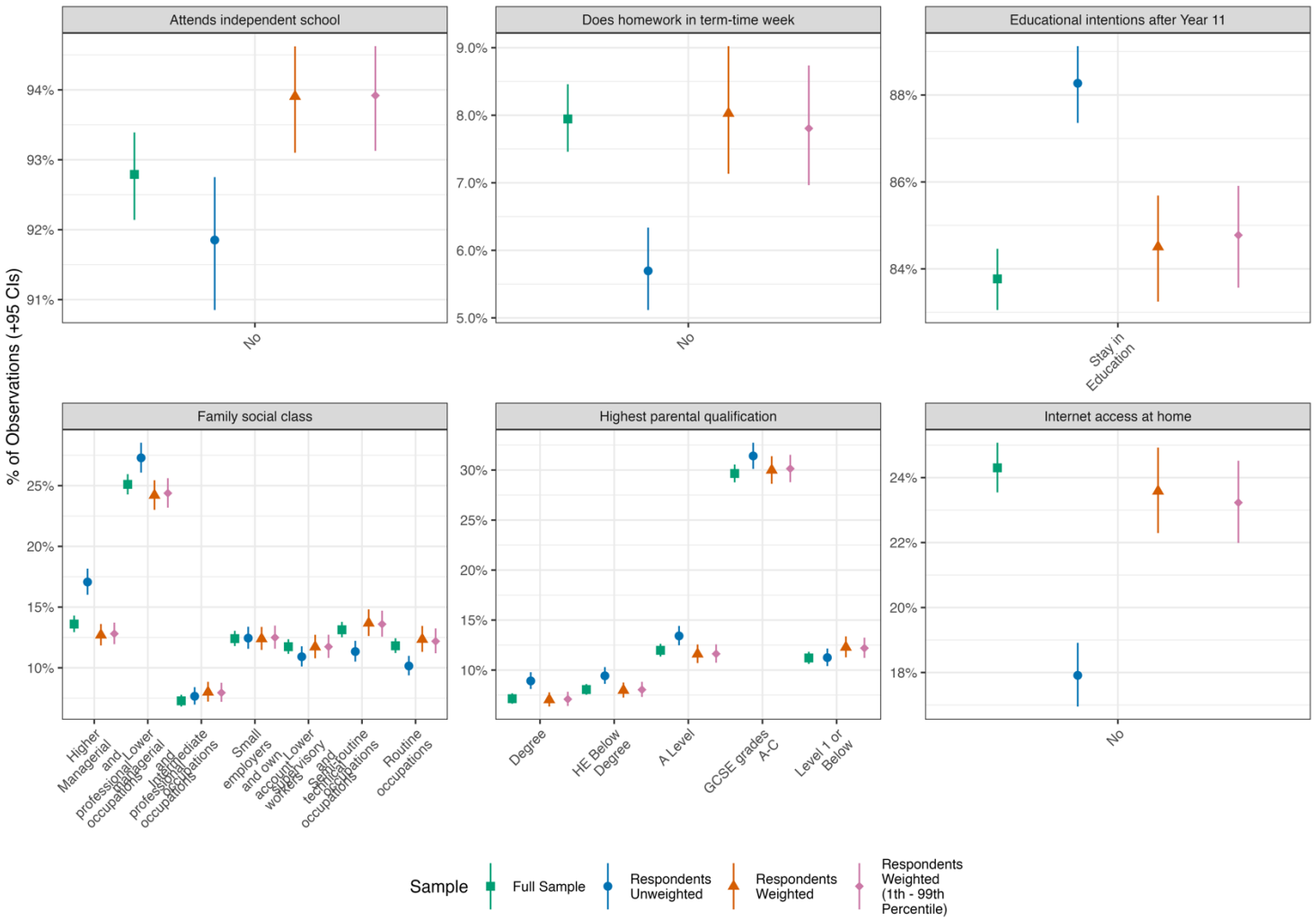
For the boost sample (variables measured in Sweep 4), initial distributions were recovered for sex, studying A-Levels and number of GCSEs being studied (Appendix Figures 8-10). Bias was reduced or unaffected in many other cases, and increased for mother's age, birth country, and ever drinking alcohol, among other variables, though again confidence intervals overlapped.



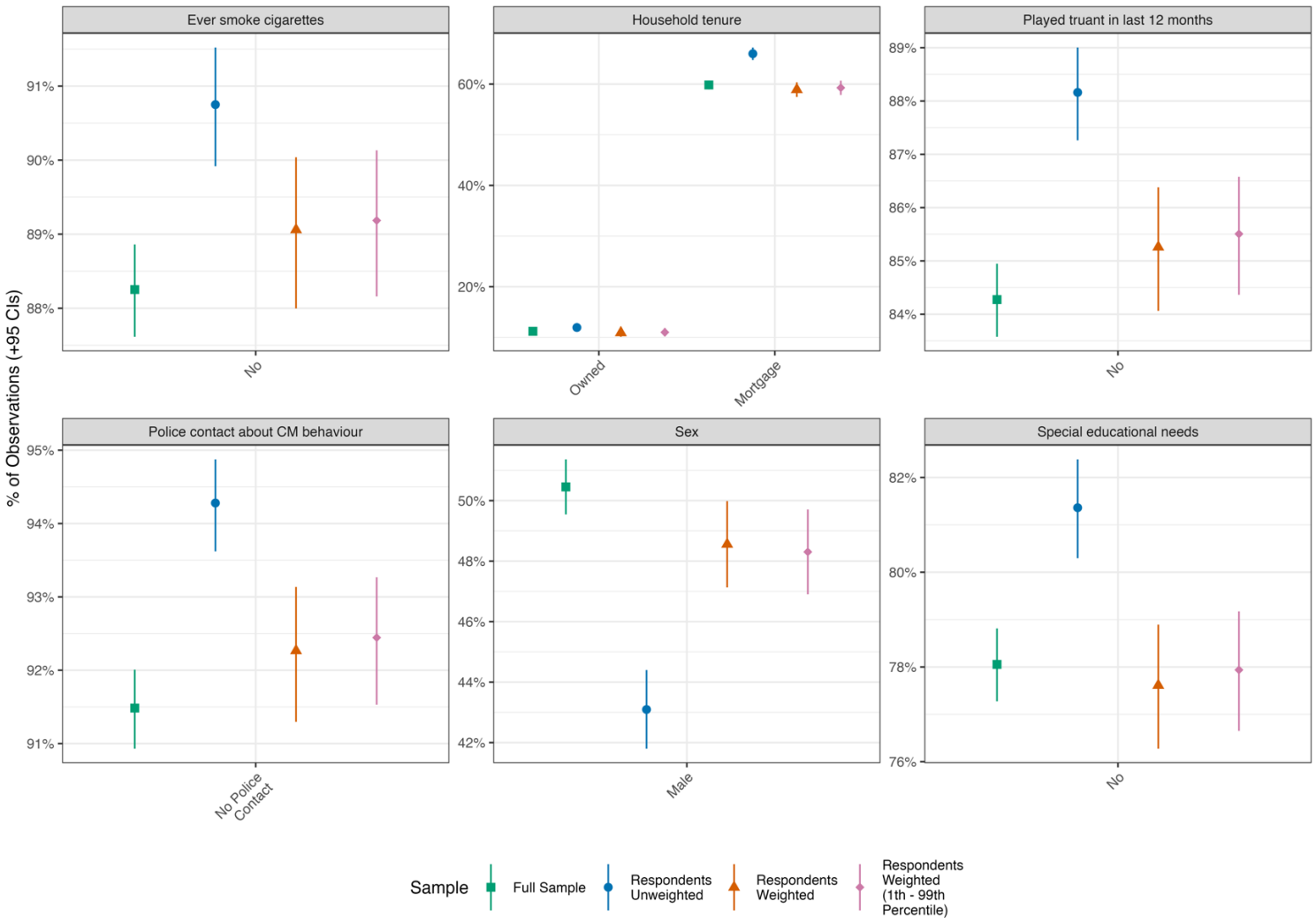
Appendix Figure 1: Means for continuous variables from Sweep 1, original cohort members – full sample and (weighted and unweighted) Sweep 9 respondents (W9FINWTALLA).



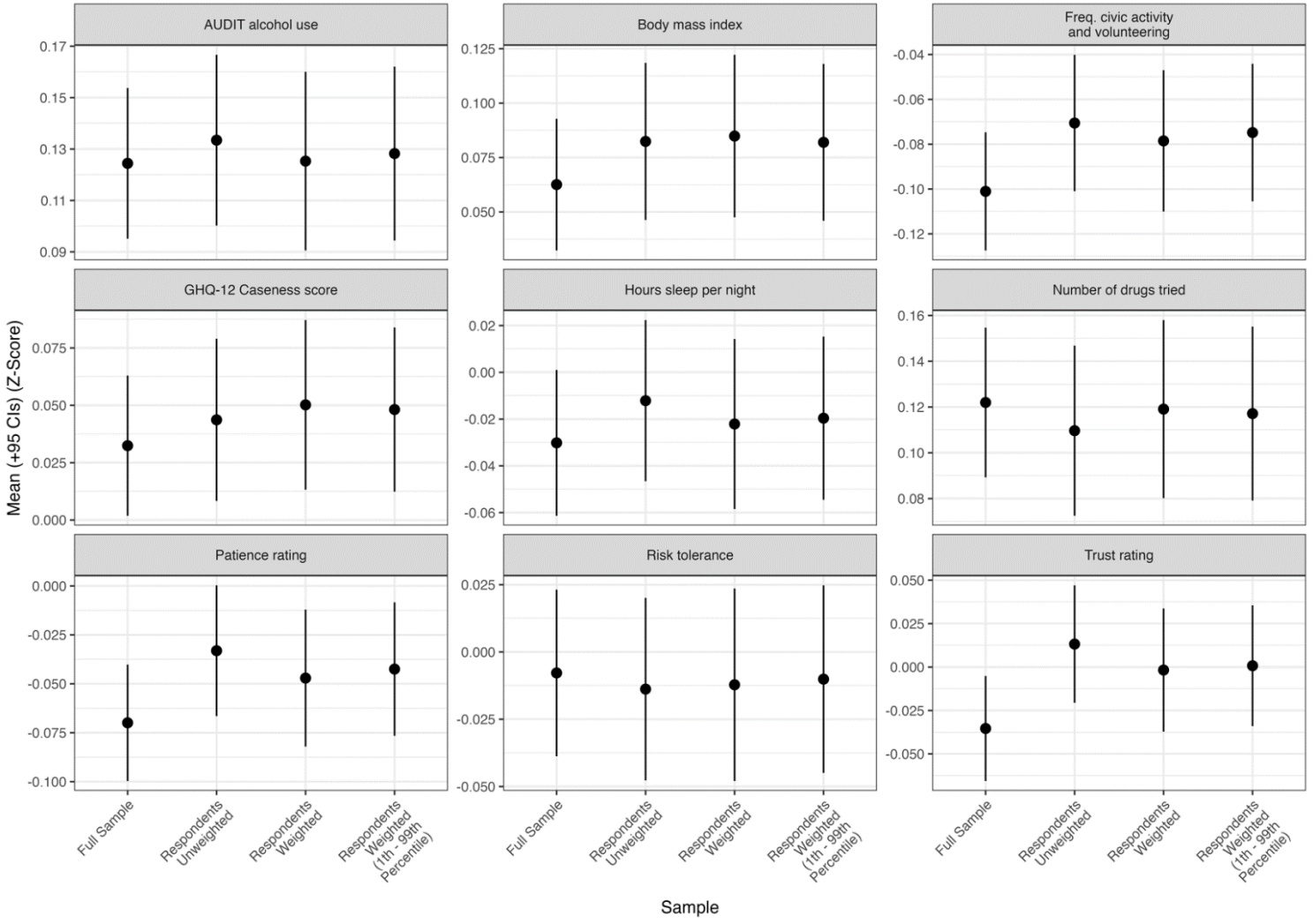
Appendix Figure 2: Proportions for values of categorical variables from Sweep 1, original cohort members – full sample and (weighted and unweighted) Sweep 9 respondents (W9FINWTALLA; also see next two figures).



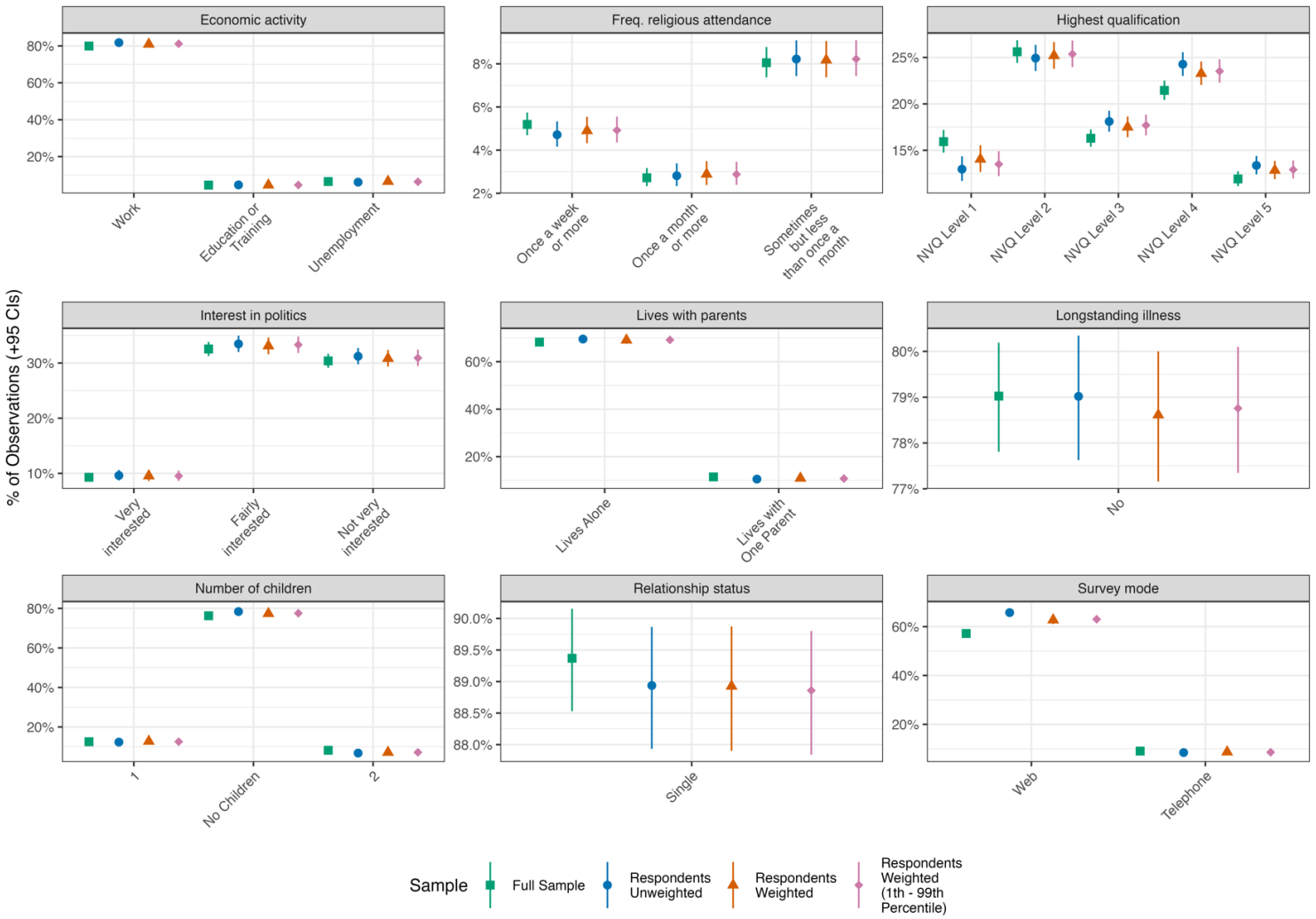
Appendix Figure 3: Proportions for values of categorical variables from Sweep 1, original cohort members – full sample and (weighted and unweighted) Sweep 9 respondents (W9FINWTALLA; also see next and previous figures).



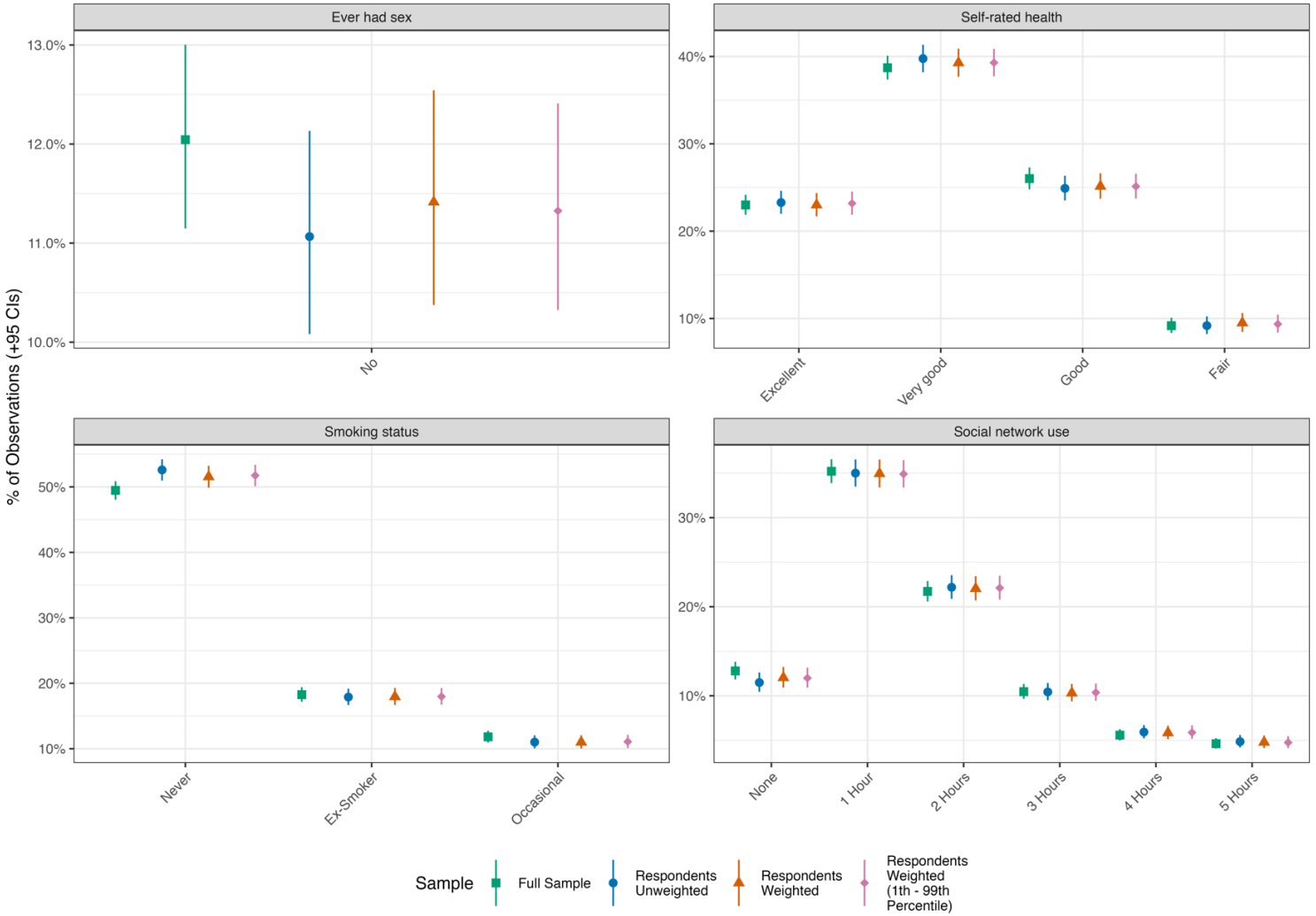
Appendix Figure 4: Proportions for values of categorical variables from Sweep 1, original cohort members – full sample and (weighted and unweighted) Sweep 9 respondents (W9FINWTALLA; also see previous two figures).



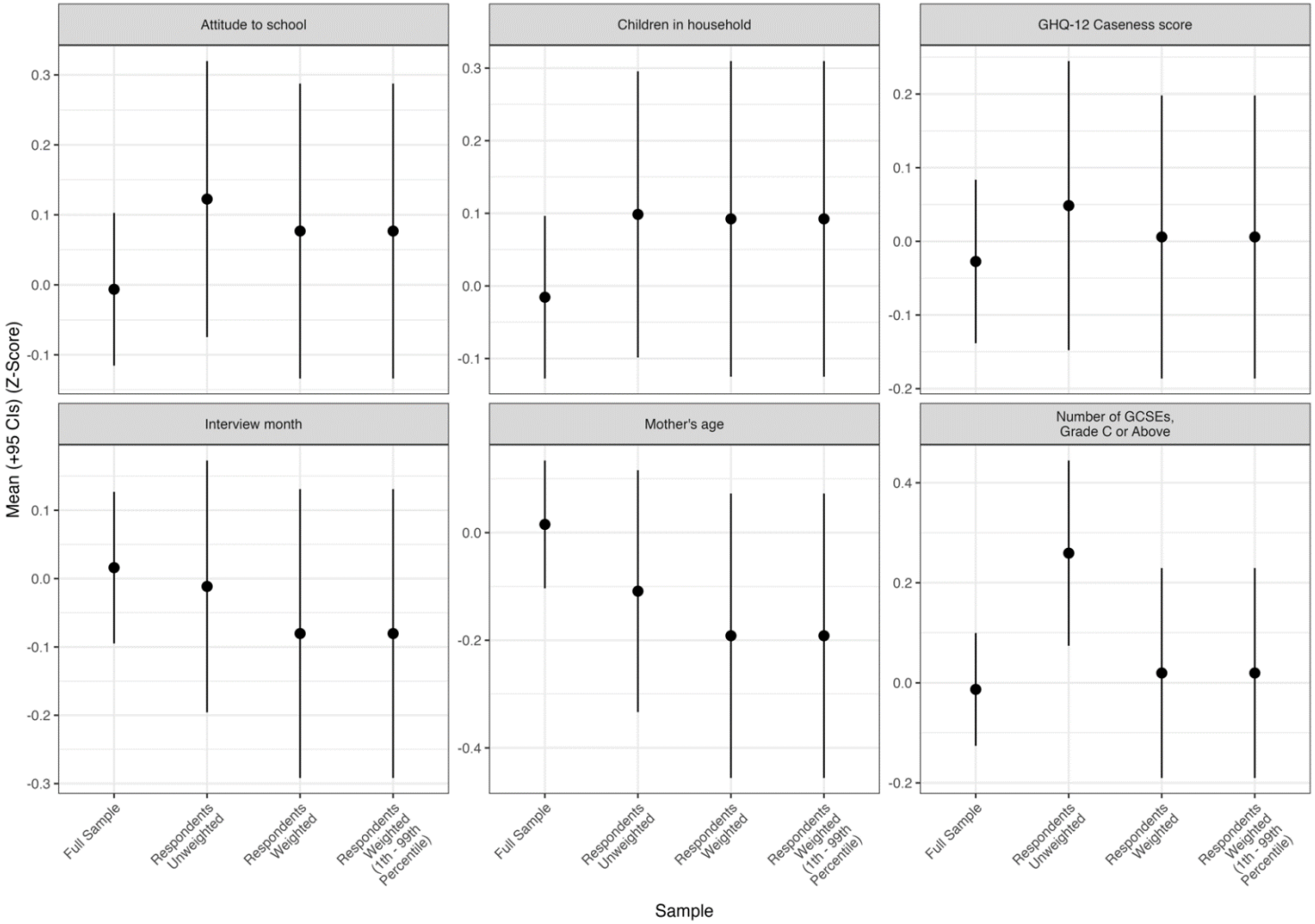
Appendix Figure 5: Means for continuous variables from Sweep 8, original cohort members – full sample and (weighted and unweighted) Sweep 9 respondents (W9FINWTALLA).



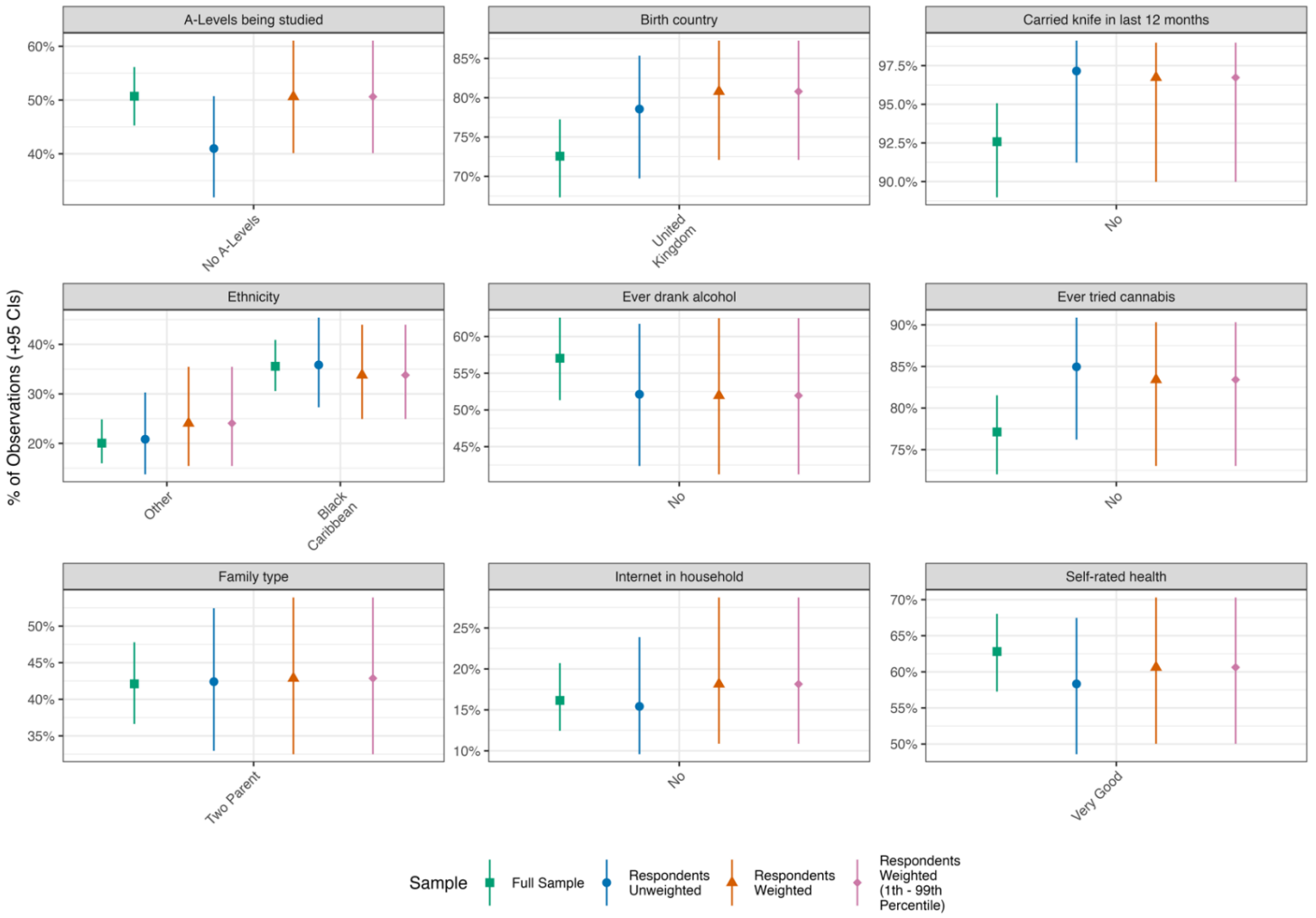
Appendix Figure 6: Proportions for values of categorical variables from Sweep 8, original cohort members – full sample and (weighted and unweighted) Sweep 9 respondents (W9FINWTALLA; also see next figure).



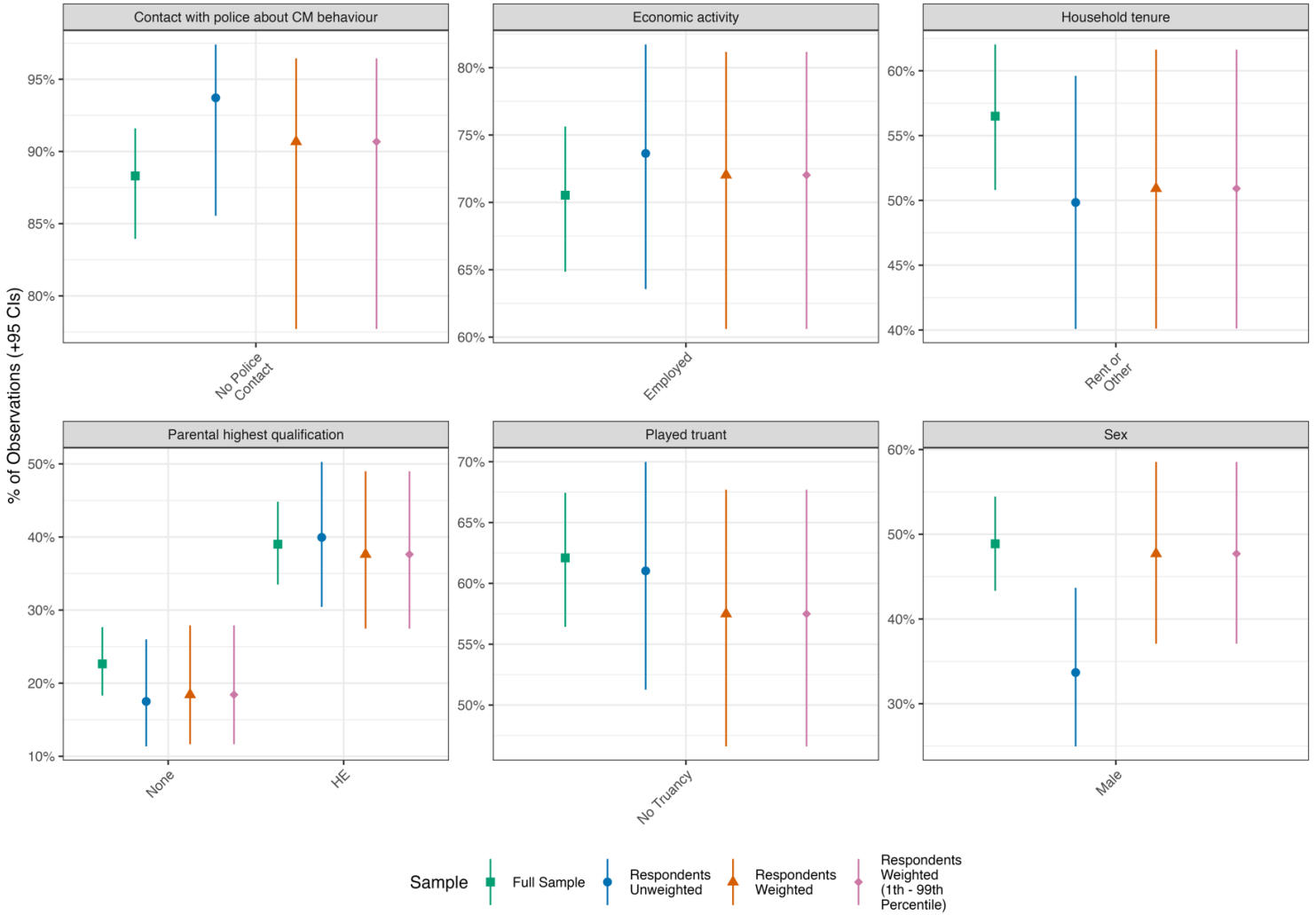
Appendix Figure 7: Proportions for values of categorical variables from Sweep 8, original cohort members – full sample and (weighted and unweighted) Sweep 9 respondents (W9FINWTALLA; also see previous figure)



Appendix Figure 8: Means for continuous variables from Sweep 4, boost cohort members – full sample and (weighted and unweighted) Sweep 9 respondents (W9FINWTALLA).



Appendix Figure 9: Proportions for values of categorical variables from Sweep 4, original cohort members – full sample and (weighted and unweighted) Sweep 9 respondents (W9FINWTALLA; also see next figure).



Appendix Figure 10: Proportions for values of categorical variables from Sweep 4, original cohort members – full sample and (weighted and unweighted) Sweep 9 respondents (W9FINWTALLA; also see previous figure)