COSMO COVID Social Mobility & Opportunities Study

Wave 1 October 2022

Data User Guide

Version 1

Tugba Adali, Jake Anders, Lisa Calderwood, Carl Cullinane, Becky Hamlyn, Jonathan Kennett, Xin Shao, Luke Taylor, David Xu

Supported by
UK Research and Innovation

Partners
The Sutton Trust
Centre for Longitudinal Studies
UCL Centre for Education Policy & Equalising Opportunities

In collaboration with
Kantar Public

Contact

Data queries: help@ukdataservice.ac.uk

Authors

Tugba Adali (CLS), Jake Anders (CEPEO), Lisa Calderwood (CLS), Carl Cullinane (Sutton Trust), Becky Hamlyn (Kantar Public), Jonathan Kennett (Kantar Public), Xin Shao (CEPEO), Luke Taylor (Kantar Public), David Xu (Kantar Public).

How to cite this guide

Adali, T., Anders, J., Calderwood, L., Cullinane, C., Hamlyn, B., Kennett, J., Shao, X., Taylor, L., Xu, David. (2022) *COVID Social Mobility & Opportunities study (COSMO): Wave 1 User Guide (Version 1)*. London: UCL Centre for Longitudinal Studies and UCL Centre for Education Policy & Equalising Opportunities.

This guide was published in August 2022 by the UCL Centre for Education Policy and Equalising Opportunities (CEPEO), UCL Centre for Longitudinal Studies (CLS), Sutton Trust, and Kantar Public.

The UCL Centre for Education Policy and Equalising Opportunities (CEPEO) is a research centre that carries out cutting-edge research focused on equalising opportunities across the life course. Its work seeks ways to improve education policy and wider practices to achieve this goal. For more information, visit www.ucl.ac.uk/ioe/cepeo.

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 manages four internationally-renowned cohort studies: the 1958 National Child Development Study, the 1970 British Cohort Study, Next Steps, and the Millennium Cohort Study. For more information, visit www.cls.ucl.ac.uk.

The Sutton Trust champions social mobility through programmes, research and policy influence. Since 1997 and under the leadership of founder Sir Peter Lampl, the Sutton Trust has worked to address low levels of social mobility in the UK. The Trust works to improve social mobility from birth to the workplace so that every young person – no matter who their parents are, what school they go to, or where they live – has the chance to succeed in life. For more information, visit www.suttontrust.com.

Kantar Public UK is an independent research agency that manages high quality cross-sectional and longitudinal research to inform the development of public policy, public service delivery and public communication on behalf of UK government and public sector organisations, academic institutions, and charities. For more information, visit www.kantarpublic.com/.

For questions and feedback about this user guide: help@ukdataservice.ac.uk

This document is available in alternative formats. Please contact the Centre for Longitudinal Studies: clsfeedback@ucl.ac.uk

Contents

1. Introduction

1.1 Background

The COVID Social Mobility & Opportunities study (COSMO) seeks to generate high-quality evidence to answer the central research question of how the COVID-19 pandemic affects socio-economic inequalities in life chances, both in terms of short-term effects on educational attainment and well-being, and long-term educational and career outcomes. To achieve this aim, a representative sample of young people who were in Year 11 in the 2021/2022 academic year across England were invited to a survey, with the intention of following them over time as they progress through the final stages of education and into the labour market. The study also included a) a survey with a parent or guardian[1] of the young person to complement the young person's data and b) a survey of the schools young people were sampled from.[2]

COSMO is carried out by a collaboration between UCL Centre for Education Policy & Equalising Opportunities (CEPEO), the UCL Centre for Longitudinal Studies (CLS), the Sutton Trust and Kantar Public. The project is further supported by key stakeholders to ensure co-production of policy-relevant evidence including: the Department for Education (DfE), the Office for Students (OfS), Administrative Data Research (ADR UK), the Education Endowment Foundation (EEF), Transforming Access and Student Outcomes in Higher Education (TASO).

This first wave of the study was funded by UKRI/ESRC as part of the COVID-19 response fund under grant ES/W001756/1. In addition, the Sutton Trust invested in an 'add on' to the main study (which we refer to as the Sutton Trust boost sample throughout this user guide), focusing on disadvantaged young people with high prior attainment.

[1] Any parent or guardian of a sampled young person was eligible for this survey. "Parents/guardians" and "parents" are used interchangeably in this guide.

[2] The data from the schools survey is not deposited due to the very low number of schools that took part. Please see section 4.4.3 for details.

COSMO was designed as a web first sequential mixed mode study both because data collection was scheduled to start at a time when COVID-19 was still a major public health concern, and because an online-first approach was thought to be suitable for young people. The data collection for Wave 1 was carried out between September 2021 and April 2022, predominantly online, but with some face-to-face interviewing. Wave 2 of the study is planned to start in October 2022 initially as online, with face-to-face fieldwork scheduled for the end of the year.

Further details on the design and implementation of the Wave 2 survey will be included in the Kantar Public Technical Report published later in 2022.

This User Guide accompanies the initial data deposit to UK Data Service. This deposit excludes data from the Sutton Trust boost sample. This will be added to the data to the deposit in due course. An application to DfE to link data from the National Pupil Database to COSMO data set has been approved. This data will be made available to researchers through the ONS Secure Research Service.

1.2 Investigators

Decisions around substantive and methodological issues on COSMO were taken by a team of investigators led by Jake Anders (CEPEO) (Principal Investigator), and including Lindsey Macmillan (CEPEO), Gill Wyness (CEPEO), Claire Crawford (CEPEO), Lisa Calderwood (CLS), Alissa Goodman (CLS), Praveetha Patalay (CLS), and Carl Cullinane (Sutton Trust).

1.3 Ethics

The study design and the tools to be used for COSMO were approved by the UCL IOE Research Ethics Committee. This application covered sampling, incentive approach, data linkage consents, participant information, privacy notice, signposting to sources of support, survey mode, questionnaires and any other relevant dimensions of the study.

2. Sampling

In this section we provide a broad overview of the target population for the study and the sampling frames used. Following on from this, a summary is provided outlining how the sample was drawn from each frame.

2.1 Target population, sampling frame and coverage

The estimation population consists of all children in England studying in Year 11 in the 2020/2021 academic year.

Two sample frames were used:

- the DfE National Pupil Database (NPD) of Year 11 children in state schools, as recorded in the Spring 2020/2021 pupil-level census[3]
- A subset of the publicly available DfE Get Information About Schools database (GIAS)[4] covering independent schools with Year 12 pupils in the 2021/2022 Academic Year

Potentially, some children will appear in both sample frames: specifically, those that moved from a state school in Year 11 to an independent school in Year 12. These respondents were identified retrospectively (via data collected in the survey questionnaire) and the weighting compensated for this (see section 6).

Those studying in very small schools were excluded from both sample frames. The total non-coverage rate among state school children was 0.8%, although it was slightly higher for children in alternative provision or special schools. The non-coverage rate among independent school children was higher: estimated[5] at 9%.

[3] The fieldwork timings (beginning in September 2021) did not allow the 2021/2022 NPD to be used for the state school sampling.

[4] https://www.get-information-schools.service.gov.uk/

[5] This is an estimate only because the number of Year 12 children in each school is not recorded in the GIAS database but is inferred by taking the total number of pupils in each school and dividing that by the number of school years covered by the school. There is

These non-covered children remain part of the estimation population and the weighting design (outlined in section 6) is designed to compensate for this non-coverage.

One other group – those children who were in an independent school in Year 11 but moved to a state school for Year 12 – are entirely uncovered. In theory, this group is part of the estimation population but, because it is missing from both sample frames, there is no way to weight the data to compensate for this non-coverage (thought to be <1%).

2.2 Sample design: State schools

In drawing the sample, we oversampled pupils from disadvantaged backgrounds (those eligible for FSM at any time in the last six years) and those from the six main minority ethnic groups (Indian, Pakistani, Bangladeshi, Black Caribbean, Black African and Mixed).

At stage one, 750 schools (Primary Sampling Units (PSUs)) were sampled using a Probability Proportionate to Size (PPS) approach. When drawing the sample, a composite size measure was used – the number of eligible students at each school weighted by their value to the study (i.e., students in groups we want to oversample will receive a larger weight). This approach allowed the disproportionate sample design to be implemented while retaining equal school-level sample sizes.

Prior to selection, schools were implicitly stratified using the following variables:

- Establishment type: Special / Alternative Provision / Other
- Admissions policy[6]: Selective / Non-selective / Missing or NA
- Region: the nine former Government Office Regions

evidence from the Independent Schools Council (ISC) that Year 12 tends to have fewer pupils than other school years, so this inferential method may well lead to an over-estimate of the number of Year 12 pupils in the school.

[6] It should be noted that this does not vary in Alternative Provision and Special schools. As such this stratification variable was only used for “other” types of establishment.

At the second stage of sampling, a stratified random sample of students was drawn from each sampled school, with sampling fractions varying between types of students. For PSUs with fewer than 50 pupils (25 schools), all Y11 pupils were selected for the study. For PSUs with more than 50 pupils (725 schools), a PPS sample of 50 pupils was drawn.

Prior to selection, pupils were implicitly stratified within each school using the following variables:

- Being eligible for FSM in the last 6 years: Yes / No
- Ethnic minority group: Indian / Bangladeshi / Pakistani / Black African / Black Caribbean / Mixed / Other
- Gender: Male / Female
- Special Educational Needs (SEN): Education, Health and Care Plan / SEN support / No Special Educational Need

The sampled schools (and their respective pupils) were then randomly allocated to original issue (460 schools) and reserve (290 schools). Following this allocation, 22,719 pupils were allocated to original issue and 14,275 to the reserve.

As will be further discussed in section 4, some reserve sampled ended up being issued into the field. Reserve sample was selected from all 290 reserve schools – a random systematic sample of 11,000 was selected from the available reserve cases available in these PSUs. In the end, there were 3,275 reserve cases that were not issued into the field.

The Sutton Trust boost sample was drawn after the main study sample was selected. The sample for the boost sample was drawn from the 460 schools selected as original issue for the main study (using the process described above). No Sutton Trust boost cases were sampled from the reserve schools.

The definition of pupils included in the boost sample was as follows:

- Eligible for FSM in last 6 years AND
- In the top 33% in the combined reading, maths, and GPS (Grammar Punctuation and Spelling) KS2 score (the score weighted as follows: maths 50%, reading 25% & GPS 25%)

In the original issue sample selected for the main study there were 22,719 pupils (within the 460 schools selected as original issue). Of these young people, 1,976 were part of Sutton Trust's population of interest (as defined above).

Within the original issue PSUs, there were a further 2,868 pupils that were eligible for the Sutton Trust boost that had not been selected for the main study. From these pupils, a further random sample of 2,000 were selected for the Sutton Trust boost (1,600 original issue and 400 reserve). As noted in section 2, all of the boost reserve sample was issued into the field.

The use of the NPD as a sampling frame for state schools was made possible through a Data Sharing Agreement[7] between UCL, Kantar Public and the DfE, following an application.

2.3 Sample design: Independent schools

A systematic random PPES (Probability Proportionate to Estimated Size) sample of 240 schools was drawn. School sampling probabilities were proportionate to the estimated number of Y12 pupils in the school.[8] There were two explicit strata: (i) independent schools (228 selected) and (ii) independent special schools (12 selected).

Before a systematic random PPES sample of schools was drawn, within each stratum schools were sorted by: region, whether they are mixed or single sex, and by whether they have boarders. This helped ensure that the sampled schools were representative of all eligible schools in terms of these factors.

Selected schools were then randomly allocated to original issue (120 schools – 114 independent and 6 independent special schools) and reserve (the

[7] DSAP number DS 00554.

[8] The GIAS database provides the number of pupils across all year groups (*NumberOfPupils*). To estimate the number of students in Y12 at each school – the total number of pupils attending the school was divided by the number of year groups (inferred from the range between the *StatutoryLowAge* and the *StatutoryHighAge* provided for each school).

remainder). As set out in section 2, the whole of the reserve ended up being issued (all 240 schools).

Cooperating schools were then asked to distribute the survey invitations to pupils and their parents/carers. Where schools had fewer than 60 pupils in the year 12 group, they were asked to invite all their pupils. For larger schools, Kantar worked with them to randomly select an appropriate number of forms to invite to the study (with the aim of inviting at least 60 pupils). For schools that did not have clearly defined forms, a suitable alternative approach was determined.

Cooperating schools provided information on the total number of forms they had in Year 12. This information was used at the weighting stage to calculate the within-school pupil sampling probability.

3. Overview of questionnaires

3.1 Development

As will be covered further below, two questionnaires were designed as part of the COSMO study: Young Person and Parent/guardian and School. These were developed over the course of May-July 2021, and were programmed into Kantar's scripting software in August 2021. A third questionnaire[9] was developed for collecting school level information from schools young people were sampled from, however data collection was not realized due to challenges in fieldwork. Please see section 4.4.3 for details.

To inform development of questionnaire content, meetings were held with various stakeholders, and input was received from researchers, governmental organisations and funders. The scientific and technical development of the questionnaires was supported by the investigators of COSMO (Lisa Calderwood, Claire Crawford, Carl Cullinane, Alissa Goodman, Lindsey Macmillan, Praveetha Patalay, Gill Wyness) led by Jake Anders, working with Kantar Public. We also gratefully acknowledge the support of Kavita Deepchand in her role as interim survey manager.

In developing the questionnaires, other relevant surveys were consulted and pre-existing questions were used or adapted where possible, to build on prior experience and ensure comparability. These surveys include, but are not limited to the Longitudinal Survey of Young People in England: Cohort 2 (LSYPE 2, also known as "Our Future"), LSYPE 1 (also known as "Next Steps"), the Millenium Cohort Study, CLS COVID-19 surveys on national longitudinal cohort studies, Science Education Tracker, and Understanding Society. A number of new questions were also developed.

[9] The questionnaire for the schools survey was developed by the COSMO team during June – August 2021. This questionnaire included questions specific to different stages of the pandemic in relation to school closures.
Because there is not data deposited for the planned schools survey, the contents of the questionnaire are not included in this guide.

To test comprehension and validity of questions, two rounds of cognitive testing were carried out with both young people and parents, focussing on a selection of proposed questions from both the young person and parent/guardian questionnaires. This informed decisions around final wording and content of these questions.

Because COSMO Wave 1 was funded by the UKRI COVID-19 rapid response fund, and needed to be in the field as quickly as possible to collect accurate information on the experiences of young people about the pandemic, the project had very tight timescales. These timescales did not allow for a pilot stage to test questionnaire flow, fieldwork processes and interview length. Therefore, a small number of informal pilot interviews were carried out by the research team using informal networks to ensure the questionnaire worked well, and to derive approximate timing estimates.

All questions for the Young Person and Parents questionnaires were designed to work in both web and face-to-face modes. For the web survey, the entire questionnaire was self-completed online. For the face-to-face survey, the more sensitive questions were administered as self-completion (CASI) which respondents completed via the interviewer's tablet.

3.2 Overview of content

The overarching aim of COSMO is to provide a representative data resource to support research into how the COVID-19 pandemic affected the life chances of pupils with different characteristics, in terms of short-term effects on educational attainment and wellbeing, and long-term educational and career outcomes. The unit of analysis is young people, but as mentioned earlier, parents were interviewed as well to complement the data collected from young people, enriching the data with information on socio-economic background, and also providing direct reports of parents' experiences during the pandemic. Below the two questionnaires are summarised further.

3.3 Young People Questionnaire

As learning about disruptions to education due to the COVID-19 pandemic at a critical point of young people's lives is central to COSMO's research questions, there is extensive information on these topics in the data. The questionnaire included questions about different periods of the pandemic, covering the two major lockdowns that took place in the UK that caused schools to close (Lockdown 1: from April to July 2020, and Lockdown 3: from January to March 2021) as well as the time in between when most schools were open (September to December 2020).

To avoid overburdening young people, two sections of the questionnaire were asked to random half samples. The first random half sample (denoted by ZMODULE=1 in the dataset) received Section I: Cancelled Assessments, while the other random half sample (ZMODULE=2) received Section K: Extra-Curricular Activities Pre- and Post-Pandemic.

Two questions in Section J: Education & Career Aspirations were only asked to a specific subgroup of the sample (ZACCESS and ZAPPLY), which was a *boost sample* funded by Sutton Trust. Please refer to the sample design section for the details of this sample.

All young people were asked for their consent to link some administrative data to their records. Details of these are provided in section 3.6.

A summary of the content is provided below in Table 1. The full questionnaires, annotated with variable names, are available within this same data release and are also available on the CLS website.

Table 1: Young Person questionnaire content at COSMO Wave 1

Section	Topic
A. Introduction, verification and opening demographics	Verification of name and address from NPD sample
	Verification of school year and name of current school (if no longer at same school as in the sample file)
	Reason for changing school if no longer in Y11 school
	Sex at birth
	Gender
	Date of birth
B. Household grid	Number of household members

Section	**Topic**
	Gender of household members
	Age group of household members
	Relationship of household members to YP
C. Current status	Current status (all activities) and main activity
	Whether apprenticeship, traineeship, internship or training course is linked to education
	Level of apprenticeship working towards
	Full/part time distinction for paid work
	If in part-time work, whether would prefer full-time
	If looking for work, hours wanting to work
	If YP not in education or training and not currently looking for work, whether looked for work in past 4 weeks
	Main reason for having left full time school/college if YP not in full time education
	If not in education or work, reasons that make it difficult to work
	Main reason for having applied to apprenticeship or training
	Characteristics of current job (shift work, holiday job, etc.)
D. Qualifications studying towards	Place of study/training
	Types of academic qualifications
	Number of academic qualifications
	Types of vocational qualifications
	Number of vocational qualifications
E. Education during lockdown 1/Year 10 (April-July 2020)	Whether attended school in person at this time
	Time spent on school work (days per week and hours per day)
	Online learning provision and attendance
	Contact with teachers/tutors
	Access to devices/provision of devices by school
	Problems related to studies
F. Education during lockdown 3/Year 11 (January-March 2021)	Whether attended school in person at this time
	Time spent on school work (days per week and hours per day)
	Online learning provision and attendance
	Contact with teachers/tutors
	Access to devices/provision of devices by school
	Problems related to studies
G. Education during Year 11 when schools have been open (September-December 2020 and March-July 2021)	Whether attended school in person at this time
	Reasons for not attending school/college in person for two or more days
	Whether affected by school closures or bubble closures
	Number of days unable to attend school/college
H. Catch up	Whether school offered catch-up activities and whether young person took them up

Section	Topic
	Concerns around disruption to education due to the COVID-19 pandemic
	How young person's motivation was affected due to pandemic
I. Cancelled assessments (asked to a random half sample)	Evaluation of teacher assessments to replace GCSE grades
	Changes in post-Year 11 plans if teacher assessment grades turned out worse or better than expected
	Concerns around cancellations of GCSE exams
	Intention to resit GSCEs in November 2021 or Summer
	Number of GSCE assessments young person did at school for main subjects
J. Education and career aspirations	Perceived likelihood of applying to go to university
	Perceived likelihood of getting in to university if they apply
	Reasons for being unlikely to apply to university
	Most likely activity in two years time
	Highest level of vocational qualification eventually aimed for those doing vocational qualifications
	Future plans to do vocational qualifications for those who are not doing vocational qualifications
	Attitudes towards statements on future life (importance of having a job/career, raising a family, etc.)
	Changes in future educational and career plans because of the pandemic
	Whether young person has an idea about courses/subjects to study at university
	Participation in activities about careers advice (careers advisors, careers fairs, university open days, etc.)
	Informal careers advice (family members, teachers, friends, etc.)
	Boost sample questions: Awareness of educational access and support programs, and whether has applied to them
K. Extra-curricular activities pre and post-pandemic (asked to a random half sample)	Participation in extra-curricular activities in Year 10 and whether organised by school or outside of school: • Sports and exercise • Other clubs (arts, crafts, music, drama, etc.) • Classes associated with church/religion • Voluntary or community work • Activities that involved overnight stays (such as Duke of Edinburgh)
	Weekly frequency of participation in all activities in Year 10
	Participation in extra-curricular activities in Year 11 after schools re-opened and whether organised by school or outside of school: • Sports and exercise • Other clubs (arts, crafts, music, drama, etc.) • Classes associated with church/religion

Section	Topic
	• Voluntary or community work • Activities that involved overnight stays (such as Duke of Edinburgh)
	Weekly frequency of participation in all activities in Year 11 after schools re-opened
L. Attitudes to education (including motivation)	Locus of control
M. Health and wellbeing (CASI)	Ever had COVID-19
	Whether have had long COVID
	Whether long COVID reduced abilities to carry out day-to-day activities
	Whether asked to shield due to clinical vulnerability
	Experience of major life events since the beginning of the pandemic
	Mental health and wellbeing scales, please see section 3.5 for details (Rosenberg scale, GHQ-12, GAD2, PHQ-2)
	Life satisfaction
	Self-assessed general health
N. Friends, peers and family support (CASI)	Peer support scale
	Social Provisions scale
	Cyber harassment
	Discrimination
	Bullying
	Evaluation of school provision of support on wellbeing and mental health
	Whether YP cares for someone who is ill, disabled or elderly and in need of care
	Ethnicity (please see section 5.10 for further explanation)
O. Health Related Behaviours (CASI)	Ever smoked, number of cigarettes per day or week
	Use of electronic cigarettes
	Ever drunk alcohol, frequency of consuming alcohol in the last 12 months, whether have had 5 or more drinks in a single occasion in the last 2 months
	Ever used cannabis, or other illegal drugs
	Sleep habits (time go to bed on week nights, time go to sleep on week nights, time wake up on week days)
	Number of times per week when young person exercised to break into a sweat (lasting at least 30 minutes, typical week over the last 4 weeks)
	If young person hurt themselves on purpose in anyway in the last 12 months
	If young person ever hurt themselves on purpose to attempt to end their lives

Section	Topic
P. Linkage	Linkage consent asked to link records from: - Department for Education (DfE) - Education Endowment Foundation - Higher Education Access Tracker - Department for Work and Pensions (DWP) - Her Majesty's Revenue and Customs (HMRC)
Q. Recontact, signposts and closing screens	Updating of young person's contact details for future waves, signposting to sources of support and closing

3.4 Parent Questionnaire

The main focus of the parent/guardian questionnaire is to complement the information obtained from young people and to provide more context on household demographics. Questions included but were not limited to parents' level of education, working status throughout the pandemic, occupation, income, and ethnicity, which provide important background information on young people. A *household reference person* approach was used when collecting information for occupation coding, so that this measure would be less dependent on the responding parent/guardian. Sections on parenting and parents' attitudes to education also help to contextualise young people's experiences.

There were also questions parents' experiences over the course of the pandemic, particularly around COVID-19 related disruptions to education, home learning and tuition, as well as impacts on household finances, and family life.

Parents'/guardians' own health and wellbeing, including their COVID-19 infection and vaccination status are also covered.

A summary of the parent questionnaire is provided below. As mentioned in the previous section, the full questionnaires, annotated with variable names, are available within this same data release and are also available on the CLS website.

Table 2: Parent questionnaire content at COSMO Wave 1

Section	Topic
A. Introduction and verification checks	Verification of being a parent/guardian of named young person (young person's name, address, school year, school, date of birth, gender)
	Gender
	Age band
	Number of household members
	Relationship of parent/guardian to YP
	Whether YP lives at same address as their mother/father
	Relationship status (legal marital status and whether lives with someone as a couple)
B. Attitudes to Education	Whether parent talks about school reports/progress reviews with YP
	Parents' aspirations for YP for after Year 11
	Parental attitudes on some statements related to YP's education
	Parents' evaluation of whether YP will go to university, and reasons if not
C. Parenting, home learning, tuition & catch-up	Parenting questions: Whether parents know where their child is going, whether they set a time for them to be back by, and how close they are to them
	Whether parent or other family members have helped YP's learning during Lockdown 1, and Lockdown 3
	Number of times YP were asked by their schools to not attend due to COVID-19 related reasons
	Attitudes on home learning (level of agreement to statements on whether parents understood school expectation, whether YP were able to manage work set by school, etc.)
	Whether YP had a tutor since Year 10, and which COVID-19 relevant time periods these were used
	Evaluation of the effect of the pandemic on YP's overall academic progress
	Parents' contact with school or teachers on issues related to COVID-19
	Additional expenditure on education during the pandemic
D. Working status across the pandemic	Main status of parent before the pandemic
	Main location of work (home, office, etc.) before the pandemic
	Work history covering from before the beginning of the pandemic until survey date (each unique status and date they ended)
	A derived variable of parent's current main status
	Whether parent was classified as a key or critical worker during the pandemic
	Whether parent experienced any changes related to their working status over the course of the pandemic (whether furloughed, whether took a pay cut, etc.)
	Main economic status of parent's partner
	Whether parent's partner is working full or part time

Section	Topic
	If parent's partner if out of work for health reasons whether they have a long-term sickness or disability
	A derived variable of parent's partner's current main status
E. Parental tenure, HRP and occupational details	Whether parent or household rents or has another arrangement
	Steps to determine household reference person: Whose name the property is owned or rented, whether parent or their partner has the highest income, whether parent or their partner is older. A derived variable on who the household reference person (HRP) is.
	For the HRP: Since when they have been in their current status, if not in work whether ever worked. Details of last job for those who had a job before: whether employee or self-employed. Main job title (open text). Open text descriptions of job, and what employer/business mainly does, for occupational coding. Whether the job required special qualifications and open text descriptions of them. Whether the job entailed managerial duties or supervision of other employees, whether more than 25 people are supervised, how many people work where HRP works as an employee, and how many employees HRP has if self-employed.
F. Parental education	Highest academic qualifications
	Highets vocational qualifications
	Other parent's highest academic qualifications
	Other parent's highest vocational qualifications
G. Parental income	Sources of income for parent and parent's partner, receipt of universal credit and other benefits, receipt of additional universal credit due to circumstances (having children, a disability or health condition, etc.)
	Banded income over a year, month or week (20 bands)
H. COVID History and vaccination (CASI)	Vaccination status
	Whether had to self-isolate, and the number of times
	Ever instructed to self-isolate but decided not to
I. Pandemic impact on family life (CASI)	Effects of the pandemic on certain aspects of life (amount of sleeping, smoking,hours of work, amount of money spent, etc.) in Lockdown 1, and in Lockdown 3
	Whether the household experienced major life events since the beginning of the pandemic (loss of a job, death of someone close, moving, etc.)
J. Parent health and wellbeing (CASI)	Mental health and wellbeing scales, please see section 3.5 for details (GHQ-12, GAD2, PHQ-2)
	Life satisfaction
	Self-assessed general health
K. Disadvantage (CASI)	Comparison of current financial situation to pre-pandemic
	Whether fallen behind on rent or mortgage since the beginning of the pandemic
	Self assessment of financial situation

Section	Topic
	Issues with housing (mould, heating issues, etc.)
	Number of bedrooms
	Food poverty and who was affected by it
	Use of a food bank since the beginning of the pandemic, and frequency of using during different periods of the pandemic
	YP's eligibility to free school meals between Year 7 and Year 11
L. Closing demographics	Ethnicity
	Whether parent is born in the UK and which country
	Religion
	Type of internet connection at home
M. Contact details, signposting and closing screens	Name and contact information for parent, whether parent lives in the same address as the YP and updating of either if necessary for future waves, signposting to sources of support and closing

3.5 Scales

The COSMO Wave 1 questionnaires included several established scales which are listed below.

3.5.1 Rosenberg Self-Esteem Scale (5-items) (Young Person questionnaire)

Rosenberg, M. (1965). Society and the adolescent self-image. Princeton, NJ: Princeton University Press.

Five items were used from the Rosenberg Self-esteem scale. This short scale was previously used in the Millennium Cohort Study. The original measure is a ten item Likert-type questionnaire. Responses are made on a 4 point scale, with answers ranging from strongly agree to strongly disagree.

The scale is thought to have good reliability and validity as a tool to measure self esteem in psychology and the social sciences. It was developed using a sample of over 5000 children drawn from schools in the state of New York and has since been widely applied since.

In COSMO, young people were asked to indicate how much they agree or disagree with statements indicating self esteem with the following responses:

1. Strongly agree

2. Agree
3. Disagree
4. Strongly disagree

The usual 10 question scale has 5 positively phrased questions and 5 negatively phrased questions. Each question gets scored from 0-3 where strongly agree=3 and strongly disagree=0 for positive questions (opposite way round for negative questions). This gives a total score of 30. Because the short form in COSMO Wave 1 uses 5 positively phrased questions, data users are recommended to score out of 15.

Variable Name	Questions
W1_ZSATI_GDSF_01	On the whole, I am satisfied with myself
W1_ZSATI_GDSF_02	I feel I have a number of good qualities
W1_ZSATI_GDSF_03	I am able to do things as well as most other people
W1_ZSATI_GDSF_04	I am a person of value
W1_ZSATI_GDSF_05	I feel good about myself

3.5.2 GHQ-12 (12 items) (Young Person questionnaire and Parent questionnaire)

Goldberg D, Williams P. A user's guide to the general health questionnaire. London: Nfer-Nelson; 1988.

The General Health Questionnaire (GHQ) is 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 present.

Variable name	Question
W1_ZGHQ1	Have you recently been able to concentrate on what you’re doing?
W1_ZGHQ2	Have you recently lost much sleep over worry?
W1_ZGHQ3	Have you recently felt that you are playing a useful part in things?
W1_ZGHQ4	Have you recently felt capable of making decisions about things?
W1_ZGHQ5	Have you recently felt constantly under strain?
W1_ZGHQ6	Have you recently felt you couldn’t overcome your difficulties?
W1_ZGHQ7	Have you recently been able to enjoy your normal day to day activities?
W1_ZGHQ8	Have you recently been able to face up to your problems?
W1_ZGHQ9	Have you recently been feeling unhappy or depressed?
W1_ZGHQ10	Have you recently been losing confidence in yourself?
W1_ZGHQ11	Have you recently been thinking of yourself as a worthless person?
W1_ZGHQ12	Have you recently been feeling reasonably happy, all things considered?

The cohort member’s score on the General Health Questionnaire 12 point scale (GHQ12) is derived by summing responses to the twelve GHQ12 questions (GHQ121 to GHQ1212). This is scored according to the 0-0-1-1 method, in which the first two possible responses to each question are assigned a value of 0 and the third and fourth responses with a value of 1, resulting in a maximum possible score of 12 for this variable. A higher score on this scale indicates a greater likelihood of mental ill health.

3.5.3 GAD-2 (Generalised Anxiety Disorder 2-item) (Young Person questionnaire)

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. GAD-2 was recently used in the COVID-19 Surveys conducted by CLS on the Millennium Cohort Study, Next Steps, 1970 British Cohort Study,1958 National Child Development Study, and MRC National Survey of Health and Development.

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

Respondents are asked whether they have been bothered by problems over the last 2 weeks, 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	Question
W1_ZGAD2PHQ2_01	Feeling nervous, anxious or on edge

W1_ZGAD2PHQ2_02	Not being able to stop or control worrying

3.5.4 PHQ-2 (Patient Health Questionnaire 2-item) (Young Person questionnaire)

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. PHQ-2 was recently used in the COVID-19 Surveys conducted by CLS on the Millennium Cohort Study, Next Steps, 1970 British Cohort Study,1958 National Child Development Study, and MRC National Survey of Health and Development.

Respondents are asked whether they have been bothered by problems over the last 2 weeks, 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	Question
W1_ZGAD2PHQ2_03	Little interest or pleasure in doing things
W1_ZGAD2PHQ2_04	Feeling down, depressed or hopeless

3.5.5 People in My Life Questionnaire (3-items) (Young Person questionnaire)

Cook E, Greenberg M, Kusche C. People in my life: Attachment relationships in middle childhood. Paper presented at the Society for Research in Child Development, Indianapolis, IN. 1995

Three items were included from the 26-item Peers Attachment Scale which is part of the People in My Life Questionnaire (the questionnaire also includes a 21-item Parents Attachment Scale). The People in My Life Questionnaire is largely based on the Inventory of Parent and Peer Attachment (IPPA) by Armsden and Greenberg (1987), but edited to be more easily comprehended by younger children (Ridenour, Greenberg and Cook, 2006).

Young people were asked about how they get on with their friends and were asked to choose the response option that best describes them and their friends using the below options:

1. Never true
2. Sometimes true
3. Often true

Always true **Variable Name**	**Questions**
W1_ZPEERSUPP_01	My friends listen to what I have to say
W1_ZPEERSUPP_02	I can count on my friends to help me when I have a problem
W1_ZPEERSUPP_03	I share my thoughts and feelings with my friends

3.5.6 Short Social Provisions Scale (3-items) (Young Person questionnaire)

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. This short scale was recently used in the COVID-19 Surveys of the Millennium Cohort Study and Next Steps, conducted by CLS.

Young people 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	Questions
W1_ZSOCPROV_01	I have family and friends who help me feel safe, secure and happy
W1_ZSOCPROV_02	There is someone I trust whom I would turn to for advice if I were having problems
W1_ZSOCPROV_03	There is no one I feel close to

3.5.7 Locus of control (Young Person questionnaire)

Young people were asked how much they agree or disagree with five items to derive a variable to indicate the extent to which they believe that they have control over events in their lives, from the following responses:

1. Strongly agree
2. Agree
3. Disagree
4. Strongly disagree

Variable Name	Questions
W1_ZSCHOOLATT2_1	If someone is not a success in life, it is usually their own fault
W1_ZSCHOOLATT2_2	People like me don't have much of a chance in life
W1_ZSCHOOLATT2_3	I can pretty much decide what will happen in my life
W1_ZSCHOOLATT2_4	How well you get on in this world is mostly a matter of luck
W1_ZSCHOOLATT2_5	If you work hard at something you'll usually succeed

The cohort members' total score on the locus of control scale was derived by summing the responses to the locus of control questions to generate a total score ranging from 5 to 20. A low value of 5 to 9 indicates an internal locus of control, a score ranging between 10 and 14 indicates either a moderate internal or moderate external locus of control, and a score between 15 and 20 suggests external locus of control.

These items have previously been asked in Next Steps Age 25 survey, as well as Next Steps Waves 7, 4 and 2 (then LSYPE1).

3.6 Data Linkage

As mentioned in section 3.3, young people were asked for their consent to link administrative data to their survey data, held by a variety of organisations:

- Education records, held by the DfE, including the National Pupil Database (NPD) and Individualised Learner Records (ILR) - covering achievement in school and further education as well as details about the school, college or training centre young people attended;
- Records about young people's enrolment in the National Tutoring Programme, held by the Education Endowment Foundation;
- Records covering students' progression from school into Higher Education and beyond, held by the Higher Education Access Tracker (HEAT);
- 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).

Taken together, consent to the linkage to NPD, ILR, DWP and HMRC records allows for linkage to the UK Government's combined Longitudinal Educational Outcomes (LEO) dataset, which is based on a combination of these administrative datasets. The procedures for explaining and obtaining these consents from young people were approved through the procedures set out by the UCL IOE Research Ethics Committee.

3.7.1 Data linkage consent process

When young people were invited to participate in COSMO, they were sent a leaflet, which explained that they would be asked for data linkage consent and it was emphasised that this was entirely their choice. Moreover, in the respondent facing website, there was a separate page on data linkage, where young people could access some frequently asked questions on data linkage. These made clear how the linkage process worked, which data holders they would be asked about and the purpose of data linkage. The webpage also emphasised that they may choose to consent to some and not other linkages, that they can complete the survey without consenting to any of them, and young people were also informed about issues like data retention and withdrawing their data.

As the young people were over the age of 16 at the time of the interview, there was no parental consent necessary for data linkage. However, on the website, it was emphasised that young people could want to discuss these with their parents if they wish to do so, and parents also received a copy of the survey leaflet which outlined this process.

Within the survey, at the beginning of the consent module, young people were informed of the steps of data linkage, that information on them will be collected on an ongoing basis unless they told the study team to stop, and that they could change their permissions at any time.

The proportion of young people who consented to linkage are presented in section 4.5.

4. Fieldwork

Wave 1 fieldwork with young people and parents was conducted between 22 September 2021 and 18 April 2022, at which stage the cohort of young people was in Year 12. All fieldwork was conducted by Kantar Public, with support from NatCen during the face-to-face stage of the study.

4.1 Fieldwork strategy

It is more usual for the first stage of a large-scale longitudinal survey to be recruited using face-to-face in-home methods, as this optimises response rates and allows for longer interview lengths. However, as all face-to-face interviewing was suspended during the COVID-19 pandemic at the time of fieldwork launch, COSMO used a sequential mixed-mode design which comprised an initial online data collection phase followed by in-home interviewing once this was allowed again. An online-first approach was also thought to be a suitable (as well as cost-effective) approach for young people, as previous research with this age group indicated that push-to-web approaches with a named sampled drawn from the National Pupil Database (NPD) could produce relatively good response rates.

Throughout fieldwork, efforts were made to maximise the number of households where both the young person and a parent participated, as this provides a more complete picture of household characteristics. Where there were two parents in a household, either parent of the sampled young person was able to participate at Wave 1.

The online phase consisted of a launch mailing followed by up to 4 reminders. As the only contact information available about issued sample members was a postal address, all communication was conducted by post. The initial survey invitation comprised two separate postal mailings for each household, where the named young person and the ‘Parent/Guardian of [named young person]’ were each sent an invitation letter that also included a survey leaflet. Reminder mailings followed a similar approach although instead of re-sending the leaflet, some FAQs were included on the reverse of the letter. Using separate mailings helped maximise the response among young people although it did lead to a relatively large number of households with only a young person and no matching parent. For the reserve sample (see section 4.1.1 below) invitation letters and reminders were sent in the same envelope for all mailings in order to reduce the number of unpaired households.

The main objectives of the face-to-face stage were to both improve response rates and help improve the sample balance, with a subset of non-responders issued to the face-to-face stage based on addresses which were least well represented after the online phase. The face-to-face stage was also used to help increase the rate of complete households (i.e. to achieve an interview with the young person or parent where only one of them had completed online).

4.1.1 Modifications to fieldwork due to further Covid-restrictions in Winter 2021

The original plan was for face-to-face fieldwork to be conducted between November 2021 and March 2022. However, this plan was adapted following further unanticipated Covid-19 restrictions introduced in England in Winter 2021 which meant that in-home face-to-face fieldwork was temporarily halted again in December 2021. As an interim measure, Kantar Public carried out a 'knock to nudge' stage in February 2022 which involved interviewers knocking on doors to encourage young people and parents to complete online. This allowed utilisation of the face-to-face interviewer panel but without any requirement for them to enter people's homes. Interviewers then returned to in-home interviewing in March 2022, and the fieldwork timeline was extended until Easter 2022 to cover as much face-to-face fieldwork as possible within the overall more limited time available.

However, the final number of face-to-face interviews achieved was still much lower than originally planned due to fieldforce capacity issues during this time. Therefore, to meet sample size targets for Wave 1, a decision was made in early 2022 to issue fresh sample from a reserve sample of addresses which had been selected at the outset alongside the main sample of addresses. At reserve sample addresses, all data collection was online as it was not possible to follow up this group using face-to-face approaches given timetable and interviewer panel resourcing constraints. The reserve sample was issued to field in mid-March and the data collection period consisted of a launch mailing and two reminders covering a field period of around four-and-a-half weeks. Key field dates are covered in section 4.3 below.

4.1.2 Fieldwork among students from Independent schools

Overall, 33 independent schools agreed to participate, and following a within-school pupil selection stage (see section 2), staff contacts at these schools sent email survey invitations to Year 12 students and their parents on Kantar's behalf. Given the generation of the sample via schools, we did not have any contact details of independent school students in advance. Therefore, students in independent schools were only contacted by web, and were not included as part of the face-to-face follow up.

In most cases only one web mailing was sent to young people and parents in the independent school sample, although some independent schools were also able to send a reminder email.

4.2 Incentives

Young people and their parents were offered a voucher conditional on survey completion to the value of either £10 or £20. The value varied depending on the nature of the school the young person attended, with those attending a school with the highest rates of pupils eligible for free school meals receiving the higher amount. Parents received an incentive to the same value as the young person in their household. Differential incentivisation was used to help ensure a good representation of students and their parents from more disadvantaged backgrounds, who are typically less likely to respond to surveys.

4.3 Key fieldwork dates

A summary of key fieldwork dates is provided below:

Table 3. Key fieldwork dates

Fieldwork phase	Sample subgroup	Dates
Original issue sample		
Initial web launch	All original issued sample	22 September
Web reminder 1	All remaining non-responders from original sample	8 October
Web reminder 2	All remaining non-responders from original sample	20 October
Initial F2F stage (halted early due to further Covid restrictions)	Non-responders selected for F2F stage	10 November - 10 December 2021
Reminder 3 mop-up web mailing	a) Remaining non-responders from original sample not issued to F2F b) Any cases initially allocated to F2F but which had not been contacted by F2F by this stage due to covid disruptions	a) 13 December b) 21 December
Knock-to nudge stage	All remaining non-responders from the F2F stage plus some additional YP and parents in new 'unpaired' households created from the previous mop-up web mailing	3 February – 4 March 2022
Return to full face-to-face	A subset of remaining non-responders from the F2F allocated sample above (we were not able to cover all addresses originally allocated to F2F due to covid-related capacity issues)	1 March - 18 April 2022
Reminder 4 final mop up web reminder	All remaining unpaired parents/YP in households which were due to be contacted F2F but did not end up being allocated to F2F	8 April 2022

Reserve sample		
Initial web launch	All reserve issued sample	17 March
Reminder 1	All remaining non-responders from reserve sample	29 March
Reminder 2	All remaining non-responders from reserve sample	7 April
All samples		
Fieldwork close	All samples	18 April 2022

4.4. Survey response

4.4.1 Achieved sample sizes - main and boost

The deposited dataset for the main sample included 9330 main sample cases which included data from a matching young person and one of their parents and a further 3498 main sample cases where we only had a young person and no matching parent interview (n=12,828 in total). Within the main sample, unmatched parent interviews, without any data from young people (n=1,602), are included in the deposited dataset but are not weighted for analysis. Young people without a matching parent (n=3,498) are still included in the dataset; however further attempts will be made to recruit a matching parent at Wave 2.

The achieved totals are provided below, split by NPD and independent school samples, and by mode. Although this dataset only includes data relating to the main sample, we have included additional figures relating to the Sutton Trust boost sample for reference (as this data will be deposited at a later date).

The original achieved field numbers were a little higher than this, but some cases were removed as part of quality assurance checks (see section 5.2 for details).

Table 4. Achieved sample size

	NPD sample				Independent school sample (main sample)	Total Main sample	Total Main sample + Boost
	Main		Boost				
	Web	F2F	Web	F2F	Web		
All young people	11,851	303	920	39	674	12,828	13,787
All parents	10,113	456	737	62	363	10,932	11,731
All young people in complete households	8,869	255	688	33	206	9,330	10,051
All parents in complete households	8,712	412	665	56	206	9,330	10,051

Response rate for the NPD main sample

The issued sample and response rates at each wave are shown below for the total NPD main sample and are also broken down by original and reserve sample. As can be seen, response rates were much higher for the original sample based on the more intensive fieldwork contact (including face-to-face) and a longer field period.

For young people, response was defined as having completed up to the end of the self-completion module and also completed the question asking for consent to linkage to DfE data (question ZYPCONDFE in Section P of the questionnaire). For parents, response was defined as having completed up to the start of the closing demographics module (question XETHNIC in Section L of the questionnaire). More information on question content is provided in section 3.

As we used a disproportionate sampling design (see section 2) and the unweighted response rates are affected by this sample design, we have also included the design-weighted response in the final column.

Table 5. Response rates for the NPD sample

	Issued sample	Achieved sample	Response rate	Response rate – (design weighted)
Total NPD main sample				
Young people	33,719	12,154	36.0%	36.9%
Parents	33,719	10,569	31.3%	32.2%
Complete household	33,719	9,124	27.1%	27.7%
Original NPD main sample only				
Young people	22,719	9,341	41.1%	42.1%
Parents	22,719	7,842	34.5%	35.6%
Complete household	22,719	6,932	30.5%	31.3%
Reserve NPD main sample only				
Young people	11,000	2,813	25.6%	26.3%
Parents	11,000	2,727	24.8%	25.6%
Complete household	11,000	2,192	19.9%	20.5%

Response rates for the NPD boost sample

The response rates, and design weighted response rates, for the boost sample are documented in the table below.

Table 6. Response rates for the NPD boost sample

	Issued sample	Achieved sample	Response rate	Response rate – (design weighted)
Total NPD boost sample				
Young people	2,000	959	48.0%	45.5%
Parents	2,000	799	40.0%	38.4%
Complete household	2,000	721	36.1%	33.6%
Original NPD boost sample only				
Young people	1,600	832	52.0%	50.0%
Parents	1,600	675	42.2%	40.4%
Complete household	1,600	620	38.8%	36.3%
Reserve NPD boost sample only				
Young people	400	127	31.8%	28.7%
Parents	400	124	31.0%	30.8%
Complete household	400	101	25.3%	23.4%

Response rates for the NPD and boost sample combined

The response rates, and design weighted response rates, for the main sample boost sample combined are documented in the table below.

Table 7. Response rates for the NPD and boost sample combined

	Issued sample	Achieved sample	Response rate	Response rate – (design weighted)
Total NPD sample (main and boost)				
Young people	35,719	13,113	36.7%	36.9%
Parents	35,719	11,368	31.8%	32.3%
Complete household	35,719	9,845	27.6%	27.8%
Original NPD sample (main and boost) only				
Young people	24,319	10,173	41.8%	42.2%
Parents	24,319	8,517	35.0%	35.6%
Complete household	24,319	7,552	31.1%	31.4%
Reserve NPD sample (main and boost) only				
Young people	11,400	2,940	25.8%	26.2%
Parents	11,400	2,851	25.0%	25.6%
Complete household	11,400	2,293	20.1%	20.5%

4.4.2 Response rates for the independent school sample

Given the different methods of sampling and fieldwork, the response rate among independent school students and their parents was considerably lower compared with students sampled via the NPD.

For the independent school sample, the response rate needs to be calculated in three stages: school level response rate; within-school response rate; and overall response rate.

School level response rate

240 schools were issued into field:

- For complete households we received survey responses from 32 schools (school response rate = 13.3%)
- For young people we received survey responses from 33 schools (school response rate = 13.8%)
- For parents we received survey responses from 32 schools (school response rate = 13.3%)

The number of participating schools differs for complete households and parents vs young people as one independent school only issued survey invitations to young people.

Within-school response rates

Based on information provided by the 32 schools that sent invitations to both parents and young people, there were 1972 pupils in the forms that were sampled.

- 206 of these responded as complete households in our final dataset (estimated household within-school response rate = 10.4%)

Based on information provided by the 33 schools that sent invitations to young people, there were 2005 pupils in the forms that were sampled.

- 674 of these responded (estimated young person within-school response rate = 33.6%)

Based on information provided by the 32 schools that sent invitations to parents, there were 1972 pupils (and parents) in the forms that were sampled.

- 363 of these responded (estimated parent within-school response rate = 18.4%)

It should be noted that these response rates are estimated as we had to rely on information provided by schools about the number of pupils in the selected forms, and we had no direct confirmation that schools sent out the exact number of invitations to pupils and their parents as we would expect based on the form selection.

Overall independent school response rate

The overall independent school response rate can be calculated by multiplying the school-level and pupil-level response rate, which produced overall response rates as below.

Pupil response rate: 13.8% x 33.6% =4.6%

Parent response rate: 13.3% x 18.4% =2.4%

Household response rate: 13.3% x 10.4%=1.4%

It is worth noting that the very low overall response rate was predominantly due to low levels of co-operation at the school level, rather than at the pupil and parent level.

4.4.3. Schools survey

As briefly mentioned in sections 1 and 3, a school staff survey was originally planned as part of COSMO. The intention had been to recruit a sample of staff members with good knowledge of the Year 11 group in the 2020-2021 academic year drawn from the same schools where young people (and their parents) were sampled to participate in the main part of the study.

However, despite significant attempts to recruit teachers to the survey using both telephone and web-based approaches, the fieldwork for this survey was adversely affected by school staff shortages and increased workload in schools during the pandemic, which led to a decision to drop this element of the research programme. The data from this component of COSMO will not be deposited for this reason, and therefore further details fall beyond the scope of this guide.

4.5 Data linkage consent rates

As discussed in section 3.6, young people were asked for their consent to link administrative data to their survey data, held four different organisations.

Consent rates and numbers of young people in the main sample, and the main and boost sample who consented to linkage are provided in the tables below.

Table 8. Linkage consent rates – main sample only

	Total no. young people in survey dataset (main)	**Total number who consented to linkage**	**Consent rate**
Department for Education	12,828	9,407	73.3%
Education Endowment Foundation	12,828	8,999	70.2%
Higher Education Access Tracker	12,828	8,834	68.9%
Department for Work and Pensions (DWP)	12,828	8,500	66.3%
Her Majesty’s Revenue and Customs (HMRC)	12,828	8,323	64.9%

Table 9. Linkage consent rates – main and boost sample

	Total no. young people in survey dataset (main and boost)	Total number who consented to linkage	Consent rate
Department for Education	13,787	10,138	73.5%
Education Endowment Foundation	13,787	9,688	70.3%
Higher Education Access Tracker	13,787	9,523	69.1%
Department for Work and Pensions (DWP)	13,787	9,138	66.3%
Her Majesty's Revenue and Customs (HMRC)	13,787	8,947	64.9%

As might be expected, consent rates for linkages were higher when interviews were conducted face-to-face. For example, for DfE linkage, the consent rate was 87.4% face-to-face compared with 77.0% online.

5. Survey Data

5.1 Summary of data

The survey data is available in two datafiles:

Young person data:

Interviews are classed as complete if all sections of the questionnaire are completed (up to the end of Section P, including ZYPCONHMRC) and as usable partial interviews if the questionnaire is completed up to the beginning of linkage questions in Section P (including ZYPCONDFE).

Table 10. Breakdown of young person interviews by type of sample and completion status

	Fully completed	Partial Complete - useable	Total
Core sample - original	9,322	19	9,341
Core sample - reserve	2,811	2	2,813
Independent school sample	669	5	674
Total: Core and independent school	12,802	26	12,828
Boost sample - original	832	0	832
Boost sample - reserve	127	0	127
Total	13,761	26	13,787

Parent data:

Interviews are classed as complete if all sections of the questionnaire are completed (up to the end of Section L, including XBRBAND), and as usable partial interviews if the questionnaire is completed until the end of the self-completion part, up to the beginning of Section L, including XETHNIC).

Table 11. Breakdown of parent interviews by type of sample and completion status

	Fully completed	**Partial Complete - useable**	**Total**
Core sample - original	7,835	7	7,842
Core sample - reserve	2,724	3	2,727
Independent school sample	362	1	363
Total: Core and independent school	10,921	11	10,932
Boost sample - original	675	0	675
Boost sample - reserve	123	1	124
Total	11,719	12	11,731

The young person is the primary cohort member so any parent interviews with no matching young person interview are not treated as part of the analytical sample. As such in the parent data file only those with a matching young person interview are weighted.

Table 12. Breakdown of interviews by type of sample and young person/parent interview matching status

	Parent in matched household	Parent not in matched household
Core sample - original	6,932	910
Core sample - reserve	2,192	535
Independent school sample	206	157
Total: Core and independent school	9,330	1,602
Boost sample - original	620	55
Boost sample - reserve	101	23
Total	10,051	1,680

Note: the boost sample is not included in the data deposit, and will be added later

These are the total number of usable interviews after removals due to data quality and validity checks, so they differ from the final fieldwork reports which include all achieved interviews.

5.2 Quality checks

The quality checks were done in 2 stages.

Stage 1: Removal of non-valid cases

The files were first cleaned to remove non-valid cases as follows:

	Description	Exclusion criteria
Unusable partials	Did not reach the completion threshold	Exclude all
Duplicates	For example, for the NPD sample, if completed on both CAWI and CAPI. And, for the independent sample, if this was completed more than once.	In these situations, we removed the least completed interview, or if the same completion status, we removed the later interview.
Wrong year	Flag if the young person birthday is not in June 2004 to end of October 2005 (this allows some buffer around the expected 1 September 2004 – 31 August 2005 school year) and school year (X/ZSYCheck) is not Year 12. For independent schools a check on form name was done for those with out-of-range birthdays.	If an interview was flagged with this status, a manual check was done across both parent/young person interviews (where we had both) to reach a decision about whether this looked to be a valid case. For independent schools, a check on form names was done to ensure we only kept cases within sampled forms.

Stage 2: Quality assurance of cases to identify those which indicate poor quality data

Based on the remaining valid cases we then assessed the data across a range of quality flags including interview length, straight-lined across attitude batteries, and repeatedly picking only one option across multi-coded questions.

Based on examining the distribution of interview lengths we decided to flag all cases where the interview length was < 0.25 * median interview length. The interview lengths for CAWI and CAPI interviews were assessed together.

For young persons, the median interview length is 39.3 minutes and < 0.25 of the median is <9.82 mins.

For parents, the median interview length is 32.4 minutes and < 0.25 of the median is < 8.1 mins.

	Description	**Exclusion criteria**
Short interview length	Flag cases where the length of interview is shorter than ¼ of the median length	Exclude all
Other indications of speeding	Each grid question was checked to see if all answers were the same (i.e. straight-lined). Also checked to see number of answers given at each multi-response question. Flag if all answer in all grid questions are straight-lined and if only one answer at all multi-response questions.	Exclude all
Independent school young person removal	The young person was not at school in England during year 11 or young person was in a state school in year 11 and surveyed as part of the NPD sample.	Exclude all

5.3 Licencing

The parent and young person 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-toaccess/registration

The datasets can be downloaded once the End User Licence access conditions have been accepted by the user. COSMO Wave 1 data available under End User Licence exclude detailed data that present a potential risk for disclosivity. This applies to:

1) Verbatim responses to open-ended questions

2) Full SOC employment codes

3) Detailed geographic information

4) School identifiers

5) Full working history since the beginning of the pandemic in the parents' data set

6) Full household grid in the young people data set

Some of these data may be made available to users within the ONS Secure Research Service, providing additional safeguards on disclosivity risk, in due course.

Please refer to section 5.9 for information on how these data have been deidentified for inclusion under End User Licence.

5.4 Identifiers

Household identifiers

The parent and young person interviews are in separate data sets and a household serial is included so interviews from the same household can be matched across the 2 datasets. This is the variable "HHserial" which is a 6-digit serial.

Individual identifiers

Each interview was assigned an individual serial, this is the “HHserial” with “1” appended for young person interviews and “2” appended for parent interviews. This is the variable “INDserial” which is a 7-digit serial.

Matching young person and parent interview into households

Because both a young person and a parent/guardian were invited to COSMO, some work has been done to ensure that we can match these interviews as young person/parent pairs (i.e. household) during data processing. Below we explain how this was done.

Data from NPD allowed the provision of unique, named invites to young people in state schools, as well as their parents (as parents of named young person). In the questionnaire, there were verification questions to make sure the invited people were filling out the survey.

For NPD sample the parent and young person were matched by sample serial. Note, the HHserial assigned in the datafiles is not the sample serial.

However, unique invites were not possible for young people in independent schools and their parents, as invitations were done at the school level and could only be unique at the school level. Therefore matching young persons and parents as households required further effort.

For independent school young persons, matching households were established by a process of reviewing responses to verification questions. As a first step the data was cleaned and split into separate young person and parent datasets to simplify the record linkage. Afterwards, the data was grouped into candidate pairs blocked by schools. All young persons were linked to all parents within a school to create all possible candidate pairs.

To assess the probability of the candidate pair being a correct match an Expectation/Conditional Maximisation (ECM) algorithm was used.

The result then underwent a manual review to establish final parent/young person matched pairs.

5.5 Variable names

Questionnaire variables in the data files were named to match the questionnaire question name whenever possible.

The standard convention used here for the naming of multi-responses and grid variables was to add a numeric suffix to the variable name in form of “VARNAME_01”. For these suffixes we consistently used _96 for “Other”, _97 for “None of These”/”None”, _98 for “Don’t Know” and _99 for “Prefer not to say”.

For wave 1 a prefix of “W1_” was added to variable names.

5.6 Variable description

For questionnaire variables the variable labels used in the data files are based on the wording from the survey questionnaire, shortened and kept comprehendible.

For multi-response and grid variables the variable label were based on the wording of the question and response text from the questionnaire. For grids the value labels used were also taken from the wording from the survey questionnaire, for multi-response variables the value labels used were No/Yes to indicate if that response was selected by the respondent.

5.7 Missing values

The missing values used in the data files are used to identify questions with no valid answer, for these there are 2 types:

1) The codes -8 and -9 are used by respondents to denote the following:

-8: Don't know

-9: Refused/Prefer not to say

These codes above, whenever they exist, were explicitly selected by respondents in the questionnaire (or communicated as such to an interviewer if CAPI).

2) The codes -1 and -2 are used for where no respondent answer was recorded:

-1: Not applicable

-2: Question not asked due to respondent answers or script

The -1 “Not Applicable” code is used if the question was intentionally not asked due to script routing. The -2 “Question not asked due to respondent answers or script” is used if questions should have been asked but wasn't asked/no data recorded. These would be cases where responses based on "Other" verbatim coding meant the script did not move down the right route, or possible script issues caused an answer to not be recorded.

There is an exception in the data sets to the use of “-1” and “-2” for useable partial interviews after the cut off points (which were XETHNIC for parents and ZYPCONDFE for young people): If the case was a usable partial interview and the codes “-1” or “-2” were required for questions after the cut off, they were set to system missing instead. As shown at the beginning of the section, this applies to a small number of useable partial interviews and a small number of variables that exist after the cut off points.

5.8 Variable order

The order of variables in the data files follow the questionnaire order as below:

- Identifier variables
- Sample information variables
- Questionnaire variables in the same order
- SIC, SOC and NSSEC variables were added in the position of the work questions.
- Para-data variables for interview device, interview time, number of interview sessions.

- Completion flag.
- Geodemographic variables
- Schools level information variables
- Weighting variables

The para-data variables included are:

- W1_DeviceDetails_kantarDevice – Device used for interviews, if multiple devices used the last used is recorded. All CAPI interviews were done on laptops.
- W1_SURVEY_SUB – The month when the interview was completed
- W1_MULTI_SESSION – Number of different sessions the interview was completed over, recorded from the number of time the survey was opened.
- W1_COMP_FLAG, Completion status of the interview. "Fully completed" – Parent: answered to XBRBAND, young person: answered to ZYPCONHMRC. "Partial Complete – useable" – Parent: answered to XETHNIC, young person: answered to ZYPCONDFE.
- Geodemographics variables included are:
- W1_Polar4_quintile – POLAR4 Quintile
- W1_Region – Region
- W1_IMD_decile – English Index of Multiple Deprivation (LSOA Decile)
- W1_IDAC_decile – English Income Deprivation Affecting Children Index (LSOA Decile)
- The school level information included are:
- W1_EstablishmentTypeGroupcode – School Establishment Type Group
- W1_AdmissionsPolicycode – School Admissions Policy
- W1_PercentageFSMQuintiles – Percentage FSM Pupils in School (Quintiles)
- W1_TrustSchoolFlagcode – Trust School Flag

5.9 Coding of disclosive information

Both data sets in this deposit have been assessed for disclosivity risk and some measures were taken to minimize the risk of identification of respondents. Below we summarize these measures.

5.9.1. Verbatim responses

The questionnaire collects some information as full verbatim answers. These have all been removed from the data files, the responses were used to either back code into existing responses or some new responses were made if there were enough verbatim answers of the same type.

Questions where new responses were added to the data based on verbatim answers were:

- W1_XRELATPAR
- W1_XASUX
- W1_XECONCHANGE
- W1_XHOMQUAL
- W1_ZSGWY
- W1_ZALEVSUB
- W1_ZASLEVSUB
- W1_ZBTECSUB
- W1_ZVCQC
- W1_ZSCHOOLMISS
- W1_ZASUX
- W1_ZCARADVINF

Responses added from coding have the note “(created from coding)” in their labels.

Employment details given in the parent survey are used to derive SIC 2020, SOC 2020 and NSSEC for either respondent or their partner. Detailed SIC and SOC codes are excluded from this deposit, but NSSEC variables were added:

- W1_XNSSEC
- W1_XPNSSEC

The NSSEC coding is based on SOC 2020 using the ONS derivation tables linked here: https://www.ons.gov.uk/file?uri=/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2020/soc2020volume3thenationalstatisticssocioeconomicclassificationnssecrebasedonthesoc2020/tables912v3.xlsx

5.9.2 Top coding/bottom coding

The higher or lower ends of the distributions of some questions were recoded to minimize the risk of identification through extreme values.

In the young person questionnaire these include:

- W1_ZDOBY, Year of birth of youth
- W1_ZHHNUM, Number of people living in household
- W1_ZHHAGE_01, Age of person 1 in household
- W1_ZHHAGE_02, Age of person 2 in household
- W1_ZHHAGE_03, Age of person 3 in household
- W1_ZHHAGE_04, Age of person 4 in household
- W1_ZHHAGE_05, Age of person 5 in household
- W1_ZHHAGE_06, Age of person 6 in household

In the parent questionnaire these include:

- W1_XDOBY, Year of birth of child
- W1_XAgePar, Age of parent
- W1_XNumPeople, Number of people living at parent's address
- W1_XWORK4AY, Year in which parent started current main working status
- W1_XPWORK4AY, Year in which parent's partner started current main working status
- W1_XINCBANDW, Weekly income bands for parent and partner

- W1_XSELFNUM, Number of separate occasions parent has had to self-isolate during the whole COVID-19 pandemic
- W1_XBEDROOM, Number of bedrooms in home

5.9.3 Sensitive information

The young person questionnaire included questions on self-harm which are deemed highly sensitive, and are left out from the dataset.

- W1_ZSelfHarm1, Whether youth has purposely hurt themselves in any way over the past 12 months
- W1_ZSelfHarm2, Whether youth has purposely hurt themselves in an attempt to end their life over the past 12 months

5.9.4 Combining response categories

Some response options were combined to reduce detail.

In the young person questionnaire these include:

- W1_ZHHREL1_01, Relationship to youth of person 1 in household
- W1_ZHHREL1_02, Relationship to youth of person 2 in household
- W1_ZHHREL1_03, Relationship to youth of person 3 in household
- W1_ZHHREL1_04, Relationship to youth of person 4 in household
- W1_ZHHREL1_05, Relationship to youth of person 5 in household
- W1_ZHHREL1_06, Relationship to youth of person 6 in household
- W1_ZYPETHNIC, Ethnic group of youth

A set of variables were also combined:

- W1_ZQUAL_04/ W1_ZQUAL_05, Qualifications currently studying for – GCSE/ IGCSE. Combined under variable W1_ZQUAL_04, and W1_ZQUAL_05 is deleted.

- W1_ZGCSENUM/ W1_ZIGCSENUM, Number of GCSEs/ IGCSEs currently studying for. Combined under variable W1_ZGCSENUM, and W1_ZIGCSENUM is deleted.
- W1_ZGCSESUB_01 – W1_ZGCSESUB_96/ W1_ZIGCSESUB_01 – W1_ZIGCSESUB_96, GCSEs/IGCSEs currently studying for. Combined under variables W1_ZGCSESUB_01 – W1_ZGCSESUB_96, and W1_ZIGCSESUB_01 – W1_ZIGCSESUB_96 are deleted.

In the parent questionnaire these include:

- W1_XGenderYP, Gender of child
- W1_XGenderPar, Gender of parent
- W1_XMarStat, Marital status of parent
- W1_XEconAcBefore, Parent's main status prior to lockdown (start of Mar 2020)
- W1_XECONACNEXT_01, Parent economic activity 1
- W1_XECONACNEXT_02, Parent economic activity 2
- W1_XECONACNEXT_03, Parent economic activity 3
- W1_XWork1, Parent's current main work status
- W1_XWorkDer, Parent's derived current work status
- W1_XPWork1, Partner's current main working status
- W1_XPWorkDer, Partner's derived working status
- W1_XTENURE, House tenure
- W1_XETHNIC, Ethnic group of parent
- W1_XCOUNTRY, Country outside UK parent was born in
- W1_XRELIGION, Religion of parent

5.9.5 Other measures to reduce detail

In addition to the changes described above other variables were removed to reduce potentially identifiable detail such as exact birthday or detailed work history, or to compliment the capping of number of people in household questions at 6 or more.

In the young person questionnaire additional deleted variables were:

- W1_ZSEX, Sex of youth at birth
- W1_ZDOBD, Day of birth of youth
- W1_ZHHGENDER_07 to W1_ZHHGENDER_15, Gender of person 7 in household to Gender of person 15 in household
- W1_ZHHAGE_07 to W1_ZHHAGE_15, Age of person 7 in household to Age of person 15 in household
- W1_ZHHREL1_07 to W1_ZHHREL1_15, Relationship to youth of person 7 in household to Relationship to youth of person 15 in household

In the parent questionnaire additional deleted variables were:

- W1_XDOBD, Day of birth of child
- W1_XECONACNEXT_04 to W1_XECONACNEXT_09, Parent economic activity 4 to Parent economic activity 9
- W1_XECONACSTOP2M_04 to W1_XECONACSTOP2M_09, Month when parent stopped economic activity 4 to Month when parent stopped economic activity 9
- W1_XECONACSTOP2Y_04 to W1_XECONACSTOP2Y_09, Year when parent stopped economic activity 4 to Year when parent stopped economic activity 9

Some non-questionnaire variables were also were also edited to reduce the amount of detail, for both the young person and parent files these were:

- W1_SURVEY_SUB, Date of survey submission
- W1_Region, Region

5.10 Data errors and inconsistencies

Users should be aware of the following error and inconsistencies in the data:

Young person questionnaire:

W1_ZFSMCHECK: The young person's free school meal status during years 7 - 11 was not asked to NPD sample until the 15th March 2021.

W1_ZYPETHNIC: The young person's ethnicity was not asked to NPD sample until the 15th March 2021.

W1_ZExCurrPost5: There was a typographical error in the school year specified in the response option 3. Instead of "3. No – didn't do this activity in Year 10 before the COVID-19 pandemic (EXCL.)", this response option should have read "3. No – didn't do this activity in Year 11 before the COVID-19 pandemic (EXCL.)", which may have caused confusion for some respondents.

Parent questionnaire:

W1_XFOODBOFT_01: The question on food bank use before the pandemic was only asked to those who reported using a foodbank since the beginning of the pandemic (W1_XFOODBANK), rather than the full sample, due to a routing error.

6. Weighting

6.1 Introduction

Weights need to be applied when conducting analysis to ensure that the sample is representative of the population and that the findings are generalisable. For this study, weights were needed for two reasons: (1) to compensate for the disproportionate sample design, and (2) to compensate for systematic non-response.

The archived dataset includes two different weight variables. The correct weight to use depends on the analysis that is being conducted:

- **W1_MainFamilyFull_weight** – should be used when analysing complete households survey data (where both the pupil and a parent in the household responded). There are weights for 9,330 households from the main survey sample.
- **W1_MainYPFull_weight** – should be used when analysing only young people's survey data (i.e., this includes data from some households where just the pupil responded to the survey). There are weights for 12,828 respondents from the main survey sample.

There were 1,602 main study households[10] where only the parent was successfully interviewed. These cases have been included in the archived dataset but have not been given a weight (the value is missing). This means that these cases will be excluded from analysis when any of the survey weights are applied.

Two further weights have been produced to analyse the survey data linked to administrative education records (from the National Pupil Database (NPD)). Separate weights are required for this analysis, as not all respondents consented to having their survey responses linked to the administrative data. These weights compensate for systematic differences in agreement rates to

[10] A further 78 for the Sutton Trust boost sample.

the linkage. These weight variables are not included in the UK Data Archive datasets, and are only included in the datasets available in the ONS Secure Research Service (SRS) where analysis with linked data will be possible.

- **W1_MainFamily_NPD_weight** - should be used when analysing survey data for complete households linked to NPD education records. There are weights for 6,896 households.
- **W1_MainYP_NPD_weight** – should be used when analysing survey data for young people only, linked to the NPD education records. There are weights for 9,385 respondents.

As outlined earlier, an additional boost sample was conducted for the Sutton Trust. The data from this boost sample will be included in an updated dataset which is planned to be published on the data archive in due course. This release will include additional weights which allow for the following analysis:

- All complete households eligible for the Sutton Boost
- All young people eligible for the Sutton Boost
- All complete households (from the main study and from the boost)
- All young people (from the main study and from the boost)

Additional weights will also be included in the ONS SRS to allow for analysis of these different samples linked to the NPD education records. The user guide will be updated to cover the details of these weights.

6.2 Approach used to derive main sample weights

A four-stage process was used to derive the main sample weights (W1_MainFamilyFull_weight and W1_MainYPFull_weight). Exactly the same process was used for both weights.

The weights included in the archived dataset are the final weights – once all stages of weighting outlined below have been completed (design weighting, non-response weighting, and calibration weighting with constraints).

This process has also been used to generate the weights that includes the Sutton Trust boost cases.

A summary of the process has been provided below:

Stage 1 – design weighting

All respondents were given a design weight equal to one divided by their sampling probability.

Respondents that were at a state school in Y11 and at an independent school in Y12 could potentially have been sampled from both sample sources used. The design weight calculated accounts for this joint selection probability.

Stage 2 – non-response modelling

All respondents were given a non-response weight equal to one divided by their estimated response probability.

For children sampled from the NPD, the estimate of response probability was a fitted value, derived from a main effects logistic regression model in which the dependent variable was a binary response indicator. The predictors included in the model were:

- Gender
- Free school meals eligibility (last 6 years)
- Ethnicity
- English as an Additional Language

- SEN provision type
- KS2 reading score (banded)
- KS2 maths score (banded)
- KS2 GPS (Grammar, Punctuation and Spelling) score (banded into terciles)
- Establishment Type (GIAS)
- Number of pupils at the school banded (GIAS)
- Percentage of population with Level 4+ qualification (Census 2011 quintiles) - based on the MSOA the school is located in
- Percentage of homes that are owned (Census 2011 quintiles) - based on the Middle Layer Super Output Area (MSOA) the school is located in
- Region (former government office region) – based on school location
- Census 2011 Output Area Classification group - based on pupil address
- ONS rural/urban classification - based on pupil address
- Income Deprivation Affecting Children Index (IDACI) quintile - based on pupil address

For the non-response weighting, missing data points were included as valid categories for variables with high levels of missing data (in particular, the KS2 variables that each had 8-9% of data missing). For other variables that had a low proportion of missing data (e.g., ethnicity) the missing data points were imputed.

For children sampled from an independent school, the estimate of response probability was a compound value based on (i) the probability that the sampled school co-operated, and (ii) the probability that the young person participated given that the school co-operated. A pair of main effect logistic regression models were used to estimate these probabilities. The predictors included in each model were the same:

- Mixed or single sex (GIAS)
- Whether school has boarders (GIAS)
- Number of pupils at the school banded (GIAS)

- Census 2011 Output Area Classification supergroup - based on school location
- Region (former government office region) - based on school location
- ONS rural/urban classification - based on school location
- Percentage of population with Level 4+ qualification (Census 2011 quintiles) - based on the MSOA the school is located in
- Percentage of homes that are owned (Census 2011 quintiles) - based on the MSOA the school is located in

For respondents who could have been sampled from both the NPD *and* the independent school sample frame, the non-response weight was derived as follows:

*(P(sampled, NPD) * P(response | sampled from NPD))*

+

*(P(sampled, independent school) * P(response | sampled from independent school))*

-

*(P(sampled, NPD) * P(response | sampled from NPD) * P(sampled, independent school) * P(response | sampled from independent school))*

This was divided by the already-calculated sampling probability to yield an estimate of response probability for these respondents.

Stage 3 - calibration

Every respondent was given a ‘base’ weight equal to one divided by the product of the sampling and estimated response probabilities.

The base-weighted respondent sample was then calibrated so that its distribution with respect to some critical variables was an *exact* match for the estimation population, *so far as this is known*.

In practice, we must use a proxy for the true estimation population, with two divisions:

- Those who were studying at a state school in Year 11 (regardless of whether sampled from NPD or from an independent school)
- Those who were studying at an independent school in *both* Years 11 and 12 (these individuals could only be sampled via an independent school)

The size of the first division of the population was equal to the number of valid records in the NPD extract of Year 11 students in Spring 2021 (= 580,278).

The calibration weight for a respondent from the first population division was equal to their base weight multiplied by a calibration factor. The iterative proportional fitting algorithm (also known as 'raking' or 'rim weighting' was then used to generate these calibration factors.

The following subclasses were included in the calibration matrix. The benchmarks used as targets for the weighting were based on the Y11 NPD Spring 2021 extract used to draw the sample:

- Size of school's Year 11:
    - Under 150 pupils
    - 150-249
    - 250+ pupils
- Type of school provision:
    - Special
    - Alternative
    - Selective Other
    - Other
- Region: 9 English regions
- FSM eligibility * SEND status:
    - FSM last 6 years & Education Health and Care (EHC) plan
    - FSM last 6 years & other SEND status
    - FSM last 6 years & no SEND status
    - No FSM last 6 years & EHC plan

  - No FSM last 6 years & other SEND status
  - No FSM last 6 years & no SEND status
- Language
  - English is primary language / not recorded
  - English is an additional language
- Sex:
  - Male
  - Female
- Ethnic group:
  - Indian
  - Bangladeshi
  - Pakistani
  - Black African
  - Black Caribbean
  - White British / no data
  - White non-British
  - Mixed / Other
- Sex * broad ethnic group:
  - Male White British
  - Male Other
  - Female White British
  - Female Other
- KS2 scores (maths / reading / GPS)
  - Upper tertile in all three
  - Upper tertile in two, middle tertile in one
  - Upper tertile in one, middle tertile in two
  - Others with at least one in upper tertile or at least two in middle tertile
  - Lower tertile in two, middle tertile in one
  - Lower tertile in all three
  - Missing data

The size of the second division of the population needs to be estimated – as there are no published official statistics for this group. There are different ways of estimating this population size, all of which are likely to be somewhat inaccurate:

> **Approach 1** – using GIAS data to estimate the population size (this is consistent with how the sample was drawn). The GIAS database provides the number of pupils across all year groups in eligible independent schools. To estimate the number of students in Y12 at each school - we divided the total number of pupils attending the school by the number of year groups. By adding this up for the 1,112 eligible independent schools, we estimate a total population of c.33,422.
>
> **Approach 2** – using published DfE KS4 data and information from the Independent Schools Council census. DfE data[11] indicates 49,597 independent school pupils took part in KS4 in 2020/2021 (which we can use as an estimate for the Y11 population size). However, available data suggests that there are fewer pupils attending Y12 of independent schools than Y11. For instance, the ISC census[12] suggests that there is a drop of c.8%pts from Y11 to Y12 (for their member schools across the UK). On this basis we might estimate that there are c.45k pupils in independent schools in England in Y12

These two approaches lead to different population size estimates. Reflecting this uncertainty, for the purpose of weighting we have used an estimated total population size of 40,000 Y12 independent school pupils. However, it is important to note that this total population includes young people that studied at state school in Y11 and that are included in the first division for the calibration stage of weighting.

The weighted first division data was used to estimate the number of pupils that attended state school in Y11 but then moved to independent for Y12

[11] https://explore-education-statistics.service.gov.uk/data-tables/permalink/e8942369-b2a3-406c-80b0-06a94a7881d6
[12] https://www.isc.co.uk/media/7496/isc_census_2021_final.pdf

(questions were included in the survey to capture this). The size of the second division could then be estimated by subtracting this figure from 40,000.

Finally, the calibration weight for a respondent from the second population division was calculated: their base weight, divided by the sum of all base weights for this division, and multiplied by the estimated population size of children studying at an independent school in both Year 11 and Year 12.

Stage 4 – constrained calibration weighting

The calibration weight divided by the design weight may be thought of as a combination of non-response weight and non-coverage weight but is mainly a non-response weight because the non-coverage level for this study was very small. We refer to this as the non-inclusion weight.

Constraining the variance of the non-inclusion weight should improve the precision of survey estimates. This can be done by trimming the non-inclusion weights (also sometimes referred as truncating). The process of trimming ensures that the minimum and maximum non-inclusion weights do not exceed (different) set values.

A respondent’s non-inclusion weight should have a theoretical lower bound equal to the response rate multiplied by the mean non-inclusion weight. There are no theoretical upper bounds for non-inclusion weights but very large, outlier values are likely to inflate the mean squared error of weighted descriptive statistics, compared to a trimmed version. It was decided that non-inclusion weights should be trimmed to be no larger than c.4 times the median value.

After trimming, the respondent sample was re-calibrated using the trimmed weights as base weights rather than the original base weights. This process was repeated until no non-inclusion weight exceeded c.4 times the median value.

With the final state school weight applied (with the stage 4 constraints), we obtained the following estimates for the number of pupils that attended a state school in Y11 and an independent school in Y12:

- W1_MainFamilyFull_weight – 4,764
- W1_MainYPFull_weight – 4,784

This left us with the following **population estimates for the population that attended independent school in both Years 11 and 12**:

- W1_MainFamilyFull_weight – 35,236 pupils (40,000 – 4,764)
- W1_MainYPFull_weight – 35,216 pupils (40,000 – 4,784)

These figures were used in the final constrained calibration weight as the estimated population size of children studying at an independent school in both Year 11 and Year 12.

6.3 Approach used to derive NPD-linked sample weights

The weights W1_MainFamily_NED_weight and W1_MainYP_NPD_weight are for use when analysing the sub-set of respondents that have agreed to NPD data linkage.[13] These weights will be available in the datasets that are deposited in the ONS SRS.

In generating these weights (W1_MainFamily_NPD_weight and W1_MainYP_NPD_weight) we used the weights previously generated (W1_MainFamilyFull_weight and W1_MainYPFull_weight) and adjusted these weights to compensate for systematic differences in consent rates for the linkage.

The approach we used was as follows:

$$\textbf{W1\_MainFamily\_NPD\_weight}_i\textbf{= W1\_MainFamilyFull\_weight}_i \textbf{ * [1/ Pr(NPD)}_i\textbf{]}$$

[13] For independent sampled pupils – these are the young people that both consented to linkage and also provided the personal information required for the linkage (full name, date of birth, school they attended in Y11).

Where:

$W1\_MainFamilyFull\_weight_i$ is the complete household weight assigned to respondent *i*; and

$Pr(NPD)_i$ is the estimated probability that respondent *i* has provided consent to NPD linkage

$$\mathbf{W1\_MainYP\_NPD\_weight_i = W1\_MainYPFull\_weight_i * [1/\ Pr(NPD)_i]}$$

Where:

$W1\_MainYPFull\_weight_i$ is the young person weight for the UCL sample that was assigned to respondent *i*; and

$Pr(NPD)_i$ is the estimated probability that respondent *i* has provided consent to NPD linkage

A logistic regression was used to estimate Pr(NPD) – the probability that a survey respondent also gave NPD linkage consent. The predictors used for this modelling are listed below.

For pupils at state school in Y11, the variables used as predictors were:

- Size of school's Year 11
- Type of school provision
- Region
- FSM eligibility * SEND status
- Language
- Ethnic group
- Sex * broad ethnic group
- KS2 scores (maths / reading / GPS)
- For pupils at independent school in Y11 and Y12, the variables used as predictors were:
- Whether school is mixed or single sex
- Whether school has boarders
- Number of pupils at the school (banded)

- Census 2011 Output Area Classification supergroup - based on school location
- Region - based on school location
- ONS rural/urban classification - based on school location
- Percentage of population with Level 4+ qualification (Census 2011 quintiles) - based on the MSOA the school is located in
- Percentage of homes that are owned (Census 2011 quintiles) - based on the MSOA the school is located in

6.4 Effectiveness of weights

To examine the effectiveness of the weights in restoring sample representativity we have compared the final weighted survey sample profiles to the benchmark population statistics (which were used when calibrating the data).

These comparisons are presented in Appendix B.

6.5 Estimation of standard errors

To ensure that standard errors are estimated correctly it is important to take into account the impact of the weighting, clustering and pre-stratification. If this is not done, the confidence intervals estimated are likely to be too narrow and there is an increased risk of Type I errors (false positives).

The variables that need to be used:

- Weight variable – as outlined in the Weighting section of this user guide (section 6), the correct weight needs to be selected for each analysis. The four weights currently available[14] are:
    - W1_MainFamilyFull_weight
    - W1_MainYPFull_weight

[14] When the Sutton Boost data is published there will be additional weights added to the dataset and this guide will be updated.

    - W1_MainFamily_NPD_weight – note this weight variable is available in the ONS SRS only
    - W1_MainYP_NPD_weight – note this weight variable is available in the ONS SRS only
- Cluster variable: W1_PSU_all
- Stratification variable*: W1_AnalysisStratum_v2

*If users run into issues when conducting sub-group analysis because of there not being two clusters in each stratum, we would suggest conducting the analysis with W1_SchoolStratum_v2. If there are further singleton stratum problems when using W1_SchoolStratum_v2, we would recommend omitting the stratification variable entirely from the survey design. While these adjustments may be necessary for standard errors to be estimated, it should be noted that they are likely to lead to slightly inflated standard error estimates.

Below we have provided exemplar code for specifying the survey design correctly in different analysis programs.

6.5.1 Stata – using the svy[15] commands

In Stata robust standard errors can be estimated using the survey commands.

Before conducting any analysis, the survey design needs to be declared for the dataset. Note that the survey design declared will need to be changed each time a different weight needs to be used (changing the text highlighted in yellow below).

```
	svyset W1_PSU_all [pweight= W1_MainFamilyFull_weight],
strata(W1_AnalysisStratum_v2)
```

[15] https://www.stata.com/manuals/svy.pdf

Subsequent commands should then be conducted using the svy prefix – e.g.,

```
svy: proportion
```

6.5.2 R – using the "survey"[16] package

First, an object specifying the survey design needs to be created. The survey design needs to reference the object in which the dataset is stored (text highlighted in green below). A different survey design object will need to be created for each weight (changing the text highlighted in yellow to reference the correct weight, and the text highlighted in grey to change the name of the object that will store each survey design).

```
library(survey)

design1 <- svydesign(id=~ W1_PSU_all,

                strata = ~ W1_AnalysisStratum_v2,

                weights=~ W1_MainFamilyFull_weight,

                data=DataObject)
```

This survey design object should then be referenced in later analysis which is conducted using the "survey" package – e.g., svymean, svyglm, etc.

6.5.3 SPSS – using the Complex Samples module[17]

A complex sample plan file needs to be saved (the file name and location need to be specified – see text highlighted in grey). Note that a separate plan file needs to be created for each weight – changing the weight variable name (highlighted in yellow) and the *.csaplan file name (highlighted in grey).

[16] https://cran.r-project.org/web/packages/survey/survey.pdf

[17] https://www.ibm.com/downloads/cas/5RWERDKG

```
CSPLAN ANALYSIS
  /PLAN FILE='\\location file should be saved\file name.csaplan'
  /PLANVARS ANALYSISWEIGHT= W1_MainFamilyFull_weight
  /SRSESTIMATOR TYPE=WR
  /PRINT PLAN
  /DESIGN STRATA=W1_AnalysisStratum_v2 CLUSTER=W1_PSU_all
  /ESTIMATOR TYPE=WR.
```

This sample plan should then be referenced when doing analysis using the Complex Samples module of SPSS – e.g. CSDESCRIPTIVES, CSTABULATE, etc.

6.6 Weighting variables in datafiles

The weighting variables in the datafiles are:

Table 13. List of weights by data deposit

		Included in UKDS data	To be included in SRS data
W1_AnalysisStratum_v2	Analysis stratum - scrambled	Y	Y
W1_SchoolStratum_v2	School stratum - scrambled	Y	Y
W1_PupilStratum_v2	Pupil stratum - scrambled	Y	Y
W1_PSU_all	PSU	Y	Y
W1_MainFamilyFull_weight	Final weight: Main Study - All Complete households	Y	Y
W1_MainYPFull_weight	Final weight: Main Study - All Young People	Y	Y
W1_MainFamily_NPD_weight	Final weight: Main Study - Complete households that consented to NPD linkage		Y
W1_MainYP_NPD_weight	Final weight: Main Study - Young People that consented to NPD linkage		Y
W1_BoostFamilyFull_weight	Final weight: Eligible for Sutton Trust Boost - All Complete households		Y
W1_BoostYPFull_weight	Final weight: Eligible for Sutton Trust Boost - All Young People		Y
W1_BoostFamily_NPD_weight	Final weight: Eligible for Sutton Trust Boost - Complete households that		Y

	consented to NPD linkage		
W1_BoostYP_NPD_weight	Final weight: Eligible for Sutton Trust Boost - Young People that consented to NPD linkage		Y
W1_AllFamilyFull_weight	Final weight: All (Main & Sutton Trust Boost) - All Complete households		Y
W1_AllYPFull_weight	Final weight: All (Main & Sutton Trust Boost) - All Young People		Y
W1_AllFamily_NPD_weight	Final weight: All (Main & Sutton Trust Boost) - Complete households that consented to NPD linkage		Y
W1_AllYP_NPD_weight	Final weight: All (Main & Sutton Trust Boost) - Young People that consented to NPD linkage		Y

7. Mode Effects

As described in sections 1 and 4, the survey involved a face-to-face phase in which a subset of web survey non-respondents were invited to take part in person.

The mode by which each respondent completed the survey is recorded in the following variables:

- For young people – W1_ZMODE
- For parents – W1_XMode

The following table shows the number of respondents that participated using each mode:

Table 14. Distribution of interviews by mode

	Total	Online	Face-to-face
Young People	12,828	12,525	303
Young People – with Boost	13,787	13,445	342
Complete households			
Parents	9,330	8,918	412
Parents – with Boost	10,051	9,583	468
Young People	9,330	9,075	255
Young People – with Boost	10,051	9,763	288

When using survey data collected using multiple modes, it is important to consider how this may affect analyses. “Mode effects” are generally taken to mean differences in observed responses to survey items which are due solely to the mode of data collection.

Attempts were made when designing the questionnaire to ensure that the online questionnaire was as similar as possible to the face-to-face approach. For instance, by using show cards for the face-to-face data collection and by making all “Don’t Know” codes explicit in both modes. Additionally, the face-to-face interviews for both parents and young person included a self-completion (CASI) section which respondents completed on their own and which included some of the sensitive items (please see section 3, Tables 1 and 2 to see content covered in the CASI sections).

Nevertheless, mode effects are unavoidable as the two approaches can never be truly identical. Some examples of why measurement may still vary between modes:

- Face-to-face interviewers can provide motivation or clarification when required; this cannot truly be replicated online
- People who would not disclose sensitive personal information or socially undesirable opinions/behaviours to an interviewer may be more willing to provide this information online

In addition, it should be noted that respondents were not randomly allocated to mode. As respondents self-selected into each mode, they are likely to differ in potentially important ways.

Without appropriate control for these (possibly unobserved) characteristics, it is not necessarily possible to determine whether an observed between-mode difference in a given variable is due to selection or truly a mode effect (or a combination of both).

8. Appendices

APPENDIX 1 – Non-response weights estimation (model outputs)

Main study full households (9,330)

Table A1. NPD model Binary logistic regression predicting whether respondents sampled from NPD participated in the study

Parameter	B	Std. Error	95% Confidence Interval		p-value
			Lower	Upper	
(Intercept)	-1.50	Er G0.24	-1.98	-1.02	0.00
Being eligible for FSM in the last 6 years? [No vs Yes]	0.04	0.03	-0.02	0.09	0.18
Ethnicity [Indian vs Other]	0.06	0.07	-0.08	0.20	0.40
Ethnicity [Pakistani vs Other]	0.08	0.07	-0.07	0.22	0.31
Ethnicity [Bangladeshi vs Other]	0.17	0.07	0.03	0.31	0.02
Ethnicity [Black Caribbean vs Other]	-0.29	0.08	-0.46	-0.13	0.00
Ethnicity [Black African vs Other]	-0.14	0.07	-0.29	0.00	0.05
Ethnicity [Mixed vs Other]	-0.12	0.07	-0.27	0.02	0.09
Ethnicity [White British vs Other]	0.00	0.06	-0.12	0.12	1.00
Ethnicity [White other vs Other]	-0.08	0.08	-0.23	0.06	0.27
English as an Additional Language [Yes vs No]	0.13	0.04	0.04	0.21	0.00
Gender [Female vs Male]	0.09	0.03	0.04	0.14	0.00
IDACI [Quintile 1 (lowest) vs Quintile 5 (highest)]	-0.06	0.06	-0.17	0.05	0.30
IDACI [Quintile 2 vs Quintile 5 (highest)]	-0.03	0.05	-0.13	0.07	0.57
IDACI [Quintile 3 vs Quintile 5 (highest)]	-0.08	0.04	-0.16	0.01	0.08
IDACI [Quintile 4 vs Quintile 5 (highest)]	-0.01	0.04	-0.09	0.06	0.73
SEN provision [EHC plan vs No SEN]	-0.23	0.09	-0.40	-0.06	0.01
SEN provision [SEN support vs No SEN]	-0.08	0.04	-0.16	0.00	0.05
KS2 reading score [Lowest tertile vs Missing data]	0.37	0.16	0.05	0.69	0.03
KS2 reading score [Middle tertile vs Missing data]	0.50	0.17	0.17	0.82	0.00
KS2 reading score [Upper tertile vs Missing data]	0.69	0.17	0.36	1.02	0.00
KS2 maths score [Lowest tertile vs Missing data]	0.21	0.18	-0.15	0.57	0.26
KS2 maths score [Middle tertile vs Missing data]	0.30	0.19	-0.07	0.66	0.11
KS2 maths score [Upper tertile vs Missing data]	0.50	0.19	0.13	0.87	0.01
KS2 Grammar Punctuation Spelling score [Lowest tertile vs Missing data]	-0.69	0.22	-1.12	-0.26	0.00
KS2 Grammar Punctuation Spelling score [Middle tertile vs Missing data]	-0.62	0.22	-1.05	-0.18	0.01
KS2 Grammar Punctuation Spelling score [Upper tertile vs Missing data]	-0.47	0.22	-0.91	-0.03	0.04

NPD model (continued)

Parameter	B	Std. Error	95% Confidence Interval		p-value
			Lower	Upper	
Output Area Classification [Ageing city dwellers vs White communities]	-0.35	0.26	-0.85	0.15	0.17
Output Area Classification [Ageing rural dwellers vs White communities]	0.13	0.16	-0.19	0.45	0.42
Output Area Classification [Ageing urban living vs White communities]	0.10	0.11	-0.12	0.32	0.37
Output Area Classification [Asian traits vs White communities]	0.11	0.12	-0.12	0.34	0.34
Output Area Classification [Aspirational techies vs White communities]	0.02	0.14	-0.26	0.29	0.91
Output Area Classification [Aspiring and affluent vs White communities]	-0.11	0.19	-0.48	0.26	0.55
Output Area Classification [Challenged Asian terraces vs White communities]	0.05	0.11	-0.16	0.26	0.65
Output Area Classification [Challenged diversity vs White communities]	0.04	0.11	-0.18	0.26	0.72
Output Area Classification [Challenged terraced workers vs White communities]	-0.11	0.12	-0.36	0.13	0.36
Output Area Classification [Comfortable cosmopolitan vs White communities]	-0.66	0.31	-1.27	-0.06	0.03
Output Area Classification [Constrained flat dwellers vs White communities]	-0.23	0.32	-0.85	0.39	0.47
Output Area Classification [Endeavouring ethnic mix vs White communities]	0.29	0.13	0.04	0.55	0.02
Output Area Classification [Ethnic dynamics vs White communities]	-0.03	0.19	-0.41	0.35	0.88
Output Area Classification [Ethnic family life vs White communities]	0.09	0.12	-0.15	0.32	0.48
Output Area Classification [Farming communities vs White communities]	0.01	0.15	-0.27	0.30	0.93
Output Area Classification [Hard pressed ageing workers vs White communities]	-0.04	0.11	-0.26	0.19	0.75
Output Area Classification [Industrious communities vs White communities]	0.07	0.11	-0.15	0.29	0.53
Output Area Classification [Inner city students vs White communities]	0.12	0.35	-0.56	0.81	0.73
Output Area Classification [Migration and churn vs White communities]	0.03	0.10	-0.18	0.23	0.80
Output Area Classification [Rented family living vs White communities]	0.07	0.10	-0.13	0.28	0.50
Output Area Classification [Rural tenants vs White communities]	0.02	0.12	-0.22	0.26	0.85
Output Area Classification [Semi-detached suburbia vs White communities]	0.06	0.11	-0.15	0.28	0.56
Output Area Classification [Students around campus vs White communities]	-0.16	0.22	-0.60	0.28	0.47
Output Area Classification [Suburban achievers vs White communities]	0.15	0.12	-0.07	0.38	0.18
Output Area Classification [Urban professionals and families vs White communities]	0.12	0.11	-0.09	0.33	0.26
Urban/rural classification [Urban conurbation vs Rural]	0.02	0.06	-0.09	0.13	0.72
Urban/rural classification [Urban city and town vs Rural]	-0.01	0.05	-0.11	0.09	0.86
Establishment type [Academies vs Special Schools]	0.30	0.20	-0.10	0.70	0.14
Establishment type [Free schools vs Special Schools]	0.25	0.22	-0.18	0.67	0.25
Establishment type [LA maintained schools vs Special Schools]	0.32	0.20	-0.07	0.72	0.11
Number of pupils [<701 vs 1,401+]	0.02	0.05	-0.07	0.11	0.70
Number of pupils [701-1,000 vs 1,401+]	0.09	0.04	0.02	0.17	0.01
Number of pupils [1,001-1,400 vs 1,401+]	0.00	0.03	-0.07	0.06	0.91

NPD model (continued)

Parameter	B	Std. Error	95% Confidence Interval		p-value
			Lower	Upper	
Percentage of population with level 4+ qualification in MSOA [Quintile 1 (highest) vs Quintile 5 (lowest)]	-0.04	0.05	-0.13	0.06	0.47
Percentage of population with level 4+ qualification in MSOA [Quintile 2 vs Quintile 5 (lowest)]	-0.02	0.05	-0.11	0.07	0.65
Percentage of population with level 4+ qualification in MSOA [Quintile 3 (highest) vs Quintile 5 (lowest)]	-0.01	0.04	-0.10	0.08	0.83
Percentage of population with level 4+ qualification in MSOA [Quintile 4 (highest) vs Quintile 5 (lowest)]	-0.10	0.04	-0.18	-0.02	0.02
Percentage of population that own their home in MSOA [Quintile 1 (highest) vs Quintile 5 (lowest)]	-0.18	0.05	-0.27	-0.08	0.00
Percentage of population that own their home in MSOA [Quintile 2 vs Quintile 5 (lowest)]	-0.03	0.05	-0.12	0.06	0.51
Percentage of population that own their home in MSOA [Quintile 3 vs Quintile 5 (lowest)]	-0.11	0.05	-0.20	-0.02	0.01
Percentage of population that own their home in MSOA [Quintile 4) vs Quintile 5 (lowest)]	0.03	0.04	-0.05	0.11	0.47
Region [East Midlands vs Yorkshire and the Humber]	0.03	0.06	-0.10	0.15	0.67
Region [East of England vs Yorkshire and the Humber]	0.01	0.06	-0.11	0.13	0.90
Region [London vs Yorkshire and the Humber]	-0.21	0.06	-0.33	-0.09	0.00
Region [North East vs Yorkshire and the Humber]	0.05	0.07	-0.09	0.19	0.51
Region [North West vs Yorkshire and the Humber]	-0.11	0.05	-0.22	0.00	0.04
Region [South East vs Yorkshire and the Humber]	-0.02	0.06	-0.13	0.10	0.76
Region [South West vs Yorkshire and the Humber]	-0.03	0.07	-0.16	0.10	0.62
Region [West Midlands vs Yorkshire and the Humber]	-0.05	0.05	-0.16	0.06	0.36

Table A2. Independent schools – school model Binary logistic regression predicting whether independent schools sampled from GIAS participated in the study

Parameter	B	Std. Error	95% Confidence Interval		p-value
			Lower	Upper	
(Intercept)	-2.89	1.99	-6.78	1.00	0.15
Output Area Classification Supergroup [Rural residents vs Suburbanites]	2.19	1.32	-0.39	4.77	0.10
Output Area Classification Supergroup [Cosmopolitan vs Suburbanites]	-0.53	1.04	-2.58	1.52	0.61
Output Area Classification Supergroup [Ethnicity central/Constrained city dwellers/Hard-pressed living/Multicultural metropolitans/Urbanites vs Suburbanites]	-0.09	0.89	-1.82	1.65	0.92
Urban/rural classification [Urban conurbation vs Rural]	1.14	1.32	-1.44	3.72	0.39
Urban/rural classification [Urban city and town vs Rural]	0.51	1.03	-1.51	2.54	0.62
Region [East Midlands vs Yorkshire and the Humber]	-2.05	1.61	-5.21	1.10	0.20
Region [East of England vs Yorkshire and the Humber]	1.18	1.02	-0.81	3.17	0.25
Region [London vs Yorkshire and the Humber]	-3.16	1.33	-5.76	-0.56	0.02
Region [North East vs Yorkshire and the Humber]	-0.55	1.75	-3.97	2.87	0.75
Region [North West vs Yorkshire and the Humber]	-1.18	1.25	-3.62	1.27	0.34
Region [South East/South West vs Yorkshire and the Humber]	-2.16	0.98	-4.08	-0.25	0.03
Region [West Midlands vs Yorkshire and the Humber]	-0.17	1.07	-2.27	1.93	0.87
Mixed or single sex [Boys vs Mixed]	-0.86	0.80	-2.43	0.72	0.29
Mixed or single sex [Girld vs Mixed]	0.75	0.59	-0.41	1.92	0.21
Whether school has boarders [Does not have boarders vs Has boarders]	0.17	0.54	-0.89	1.23	0.75
Number of pupils [<701 vs 1,201+]	-1.08	0.76	-2.57	0.41	0.16
Number of pupils [701-1,000 vs 1,201+]	-0.62	0.72	-2.02	0.79	0.39
Number of pupils [1,001-1,200 vs 1,201+]	-0.97	0.83	-2.60	0.66	0.24
Percentage of population with level 4+ qualification in MSOA [Quintile 1 (highest) vs Quintile 5 (lowest)]	2.20	1.11	0.02	4.38	0.05
Percentage of population with level 4+ qualification in MSOA [Quintile 2 vs Quintile 5 (lowest)]	3.68	1.14	1.46	5.91	0.00
Percentage of population with level 4+ qualification in MSOA [Quintile 3 (highest) vs Quintile 5 (lowest)]	2.16	1.02	0.17	4.16	0.03
Percentage of population with level 4+ qualification in MSOA [Quintile 4 (highest) vs Quintile 5 (lowest)]	1.89	1.03	-0.13	3.91	0.07
Percentage of population that own their home in MSOA [Quintile 1 (highest) vs Quintile 5 (lowest)]	-0.54	0.98	-2.46	1.37	0.58
Percentage of population that own their home in MSOA [Quintile 2 vs Quintile 5 (lowest)]	-1.42	1.00	-3.39	0.54	0.16
Percentage of population that own their home in MSOA [Quintile 3 vs Quintile 5 (lowest)]	-1.04	0.97	-2.93	0.86	0.28
Percentage of population that own their home in MSOA [Quintile 4) vs Quintile 5 (lowest)]	1.79	0.86	0.11	3.46	0.04

Table A3. Independent schools – young person model Binary logistic regression predicting whether young people sampled from cooperating independent schools participated in the study

Parameter	B	Std. Error	95% Confidence Interval		p-value
			Lower	Upper	
(Intercept)	-3.31	1.79	-6.82	0.20	0.06
Output Area Classification Supergroup [Rural residents vs Suburbanites]	3.45	1.57	0.37	6.52	0.03
Output Area Classification Supergroup [Cosmopolitan vs Suburbanites]	0.68	0.58	-0.46	1.81	0.24
Output Area Classification Supergroup [Ethnicity central/Constrained city dwellers/Hard-pressed living/Multicultural metropolitans/Urbanites vs Suburbanites]	0.76	0.48	-0.17	1.69	0.11
Urban/rural classification [Urban conurbation vs Rural]	1.88	1.47	-1.00	4.75	0.20
Urban/rural classification [Urban city and town vs Rural]	-0.27	0.83	-1.90	1.36	0.74
Region [East Midlands vs Yorkshire and the Humber]	-2.37	1.82	-5.94	1.20	0.19
Region [East of England vs Yorkshire and the Humber]	-0.20	0.58	-1.34	0.94	0.73
Region [London vs Yorkshire and the Humber]	-3.47	1.28	-5.98	-0.95	0.01
Region [North East vs Yorkshire and the Humber]	-3.74	2.02	-7.71	0.23	0.06
Region [North West vs Yorkshire and the Humber]	-3.45	1.30	-6.00	-0.90	0.01
Region [South East/South West vs Yorkshire and the Humber]	0.60	0.69	-0.74	1.95	0.38
Region [West Midlands vs Yorkshire and the Humber]	-1.53	0.90	-3.29	0.24	0.09
Mixed or single sex [Boys vs Mixed]	-1.19	0.55	-2.27	-0.11	0.03
Mixed or single sex [Girld vs Mixed]	0.43	0.35	-0.26	1.12	0.22
Whether school has boarders [Does not have boarders vs Has boarders]	-0.04	0.36	-0.74	0.67	0.92
Number of pupils [<701 vs 1,201+]	-1.24	0.48	-2.18	-0.29	0.01
Number of pupils [701-1,000 vs 1,201+]	-0.45	0.65	-1.73	0.83	0.49
Number of pupils [1,001-1,200 vs 1,201+]	-0.76	0.44	-1.63	0.10	0.08
Percentage of population with level 4+ qualification in MSOA [Quintile 1 (highest) vs Quintile 5 (lowest)]	2.77	1.06	0.70	4.84	0.01
Percentage of population with level 4+ qualification in MSOA [Quintile 2 vs Quintile 5 (lowest)]	1.71	0.87	0.01	3.41	0.05
Percentage of population with level 4+ qualification in MSOA [Quintile 3 (highest) vs Quintile 5 (lowest)]	2.84	1.45	-0.01	5.69	0.05
Percentage of population with level 4+ qualification in MSOA [Quintile 4 (highest) vs Quintile 5 (lowest)]	3.04	1.32	0.44	5.63	0.02
Percentage of population that own their home in MSOA [Quintile 1 (highest) vs Quintile 5 (lowest)]	-2.89	1.03	-4.91	-0.88	0.00
Percentage of population that own their home in MSOA [Quintile 2 vs Quintile 5 (lowest)]	-1.97	0.70	-3.33	-0.61	0.00
Percentage of population that own their home in MSOA [Quintile 3 vs Quintile 5 (lowest)]	-0.48	0.43	-1.33	0.36	0.26
Percentage of population that own their home in MSOA [Quintile 4) vs Quintile 5 (lowest)]	-0.64	0.65	-1.91	0.63	0.32

Main study Young People (12,828)

Table A4. NPD model Binary logistic regression predicting whether respondents sampled from NPD participated in the study

Parameter	B	Std. Error	95% Confidence Interval		p-value
			Lower	Upper	
(Intercept)	-1.23	0.22	-1.66	-0.79	0.00
Being eligible for FSM in the last 6 years? [No vs Yes]	0.04	0.03	-0.01	0.09	0.16
Ethnicity [Indian vs Other]	0.04	0.07	-0.09	0.18	0.51
Ethnicity [Pakistani vs Other]	0.04	0.07	-0.10	0.18	0.56
Ethnicity [Bangladeshi vs Other]	0.11	0.07	-0.03	0.24	0.12
Ethnicity [Black Caribbean vs Other]	-0.38	0.08	-0.53	-0.23	0.00
Ethnicity [Black African vs Other]	-0.05	0.07	-0.18	0.09	0.49
Ethnicity [Mixed vs Other]	-0.17	0.07	-0.31	-0.04	0.01
Ethnicity [White British vs Other]	-0.05	0.05	-0.16	0.06	0.37
Ethnicity [White other vs Other]	-0.12	0.07	-0.26	0.02	0.10
English as an Additional Language [Yes vs No]	0.09	0.04	0.01	0.16	0.03
Gender [Female vs Male]	0.17	0.02	0.12	0.21	0.00
IDACI [Quintile 1 (lowest) vs Quintile 5 (highest)]	-0.03	0.05	-0.13	0.08	0.63
IDACI [Quintile 2 vs Quintile 5 (highest)]	0.00	0.05	-0.09	0.10	0.95
IDACI [Quintile 3 vs Quintile 5 (highest)]	-0.01	0.04	-0.09	0.07	0.88
IDACI [Quintile 4 vs Quintile 5 (highest)]	0.02	0.03	-0.05	0.09	0.61
SEN provision [EHC plan vs No SEN]	-0.26	0.08	-0.41	-0.11	0.00
SEN provision [SEN support vs No SEN]	-0.10	0.04	-0.17	-0.03	0.01
KS2 reading score [Lowest tertile vs Missing data]	0.19	0.14	-0.08	0.47	0.17
KS2 reading score [Middle tertile vs Missing data]	0.33	0.14	0.05	0.61	0.02
KS2 reading score [Upper tertile vs Missing data]	0.54	0.15	0.25	0.82	0.00
KS2 maths score [Lowest tertile vs Missing data]	0.05	0.17	-0.28	0.38	0.76
KS2 maths score [Middle tertile vs Missing data]	0.16	0.17	-0.17	0.49	0.35
KS2 maths score [Upper tertile vs Missing data]	0.38	0.17	0.05	0.72	0.02
KS2 Grammar Punctuation Spelling score [Lowest tertile vs Missing data]	-0.35	0.20	-0.73	0.04	0.08
KS2 Grammar Punctuation Spelling score [Middle tertile vs Missing data]	-0.29	0.20	-0.68	0.10	0.14
KS2 Grammar Punctuation Spelling score [Upper tertile vs Missing data]	-0.12	0.20	-0.51	0.27	0.54

NPD model (continued)

Parameter	B	Std. Error	95% Confidence Interval		p-value
			Lower	Upper	
Output Area Classification [Ageing city dwellers vs White communities]	-0.31	0.23	-0.76	0.14	0.17
Output Area Classification [Ageing rural dwellers vs White communities]	0.06	0.15	-0.23	0.36	0.67
Output Area Classification [Ageing urban living vs White communities]	0.09	0.11	-0.11	0.30	0.38
Output Area Classification [Asian traits vs White communities]	0.08	0.11	-0.13	0.29	0.43
Output Area Classification [Aspirational techies vs White communities]	0.01	0.13	-0.25	0.27	0.94
Output Area Classification [Aspiring and affluent vs White communities]	-0.11	0.17	-0.45	0.23	0.53
Output Area Classification [Challenged Asian terraces vs White communities]	0.01	0.10	-0.19	0.21	0.91
Output Area Classification [Challenged diversity vs White communities]	0.03	0.10	-0.17	0.23	0.78
Output Area Classification [Challenged terraced workers vs White communities]	-0.03	0.11	-0.25	0.20	0.82
Output Area Classification [Comfortable cosmopolitan vs White communities]	-0.86	0.29	-1.42	-0.30	0.00
Output Area Classification [Constrained flat dwellers vs White communities]	-0.29	0.29	-0.86	0.28	0.32
Output Area Classification [Endeavouring ethnic mix vs White communities]	0.27	0.12	0.04	0.50	0.02
Output Area Classification [Ethnic dynamics vs White communities]	-0.05	0.18	-0.40	0.30	0.78
Output Area Classification [Ethnic family life vs White communities]	0.13	0.11	-0.09	0.35	0.26
Output Area Classification [Farming communities vs White communities]	-0.03	0.14	-0.29	0.24	0.85
Output Area Classification [Hard pressed ageing workers vs White communities]	-0.04	0.11	-0.24	0.17	0.72
Output Area Classification [Industrious communities vs White communities]	0.09	0.10	-0.11	0.30	0.36
Output Area Classification [Inner city students vs White communities]	0.11	0.33	-0.53	0.75	0.74
Output Area Classification [Migration and churn vs White communities]	0.08	0.10	-0.11	0.27	0.40
Output Area Classification [Rented family living vs White communities]	0.11	0.10	-0.08	0.30	0.24
Output Area Classification [Rural tenants vs White communities]	0.05	0.11	-0.17	0.28	0.64
Output Area Classification [Semi-detached suburbia vs White communities]	0.06	0.10	-0.14	0.25	0.57
Output Area Classification [Students around campus vs White communities]	-0.23	0.21	-0.64	0.18	0.26
Output Area Classification [Suburban achievers vs White communities]	0.12	0.11	-0.09	0.33	0.26
Output Area Classification [Urban professionals and families vs White communities]	0.12	0.10	-0.08	0.31	0.23
Urban/rural classification [Urban conurbation vs Rural]	0.04	0.05	-0.07	0.14	0.48
Urban/rural classification [Urban city and town vs Rural]	-0.01	0.05	-0.10	0.08	0.87
Establishment type [Academies vs Special Schools]	0.36	0.18	0.00	0.71	0.05
Establishment type [Free schools vs Special Schools]	0.28	0.20	-0.10	0.66	0.15
Establishment type [LA maintained schools vs Special Schools]	0.38	0.18	0.02	0.74	0.04
Number of pupils [<701 vs 1,401+]	0.02	0.04	-0.06	0.11	0.56
Number of pupils [701-1,000 vs 1,401+]	0.09	0.03	0.02	0.16	0.01
Number of pupils [1,001-1,400 vs 1,401+]	0.00	0.03	-0.06	0.06	0.97

NPD model (continued)

Parameter	B	Std. Error	95% Confidence Interval		p-value
			Lower	Upper	
Percentage of population with level 4+ qualification in MSOA [Quintile 1 (highest) vs Quintile 5 (lowest)]	-0.02	0.05	-0.11	0.07	0.72
Percentage of population with level 4+ qualification in MSOA [Quintile 2 vs Quintile 5 (lowest)]	-0.01	0.04	-0.10	0.07	0.74
Percentage of population with level 4+ qualification in MSOA [Quintile 3 (highest) vs Quintile 5 (lowest)]	0.00	0.04	-0.08	0.08	1.00
Percentage of population with level 4+ qualification in MSOA [Quintile 4 (highest) vs Quintile 5 (lowest)]	-0.08	0.04	-0.16	0.00	0.04
Percentage of population that own their home in MSOA [Quintile 1 (highest) vs Quintile 5 (lowest)]	-0.18	0.05	-0.27	-0.09	0.00
Percentage of population that own their home in MSOA [Quintile 2 vs Quintile 5 (lowest)]	-0.07	0.04	-0.16	0.02	0.12
Percentage of population that own their home in MSOA [Quintile 3 vs Quintile 5 (lowest)]	-0.13	0.04	-0.22	-0.05	0.00
Percentage of population that own their home in MSOA [Quintile 4) vs Quintile 5 (lowest)]	-0.02	0.04	-0.10	0.06	0.58
Region [East Midlands vs Yorkshire and the Humber]	0.11	0.06	-0.01	0.22	0.07
Region [East of England vs Yorkshire and the Humber]	0.11	0.06	0.00	0.22	0.05
Region [London vs Yorkshire and the Humber]	-0.19	0.06	-0.30	-0.08	0.00
Region [North East vs Yorkshire and the Humber]	0.02	0.07	-0.11	0.15	0.78
Region [North West vs Yorkshire and the Humber]	-0.07	0.05	-0.17	0.03	0.16
Region [South East vs Yorkshire and the Humber]	0.04	0.05	-0.06	0.15	0.43
Region [South West vs Yorkshire and the Humber]	0.03	0.06	-0.09	0.15	0.57
Region [West Midlands vs Yorkshire and the Humber]	0.02	0.05	-0.08	0.12	0.70

Table A5. Independent schools – school model Binary logistic regression predicting whether independent schools sampled from GIAS participated in the study

Parameter	B	Std. Error	95% Confidence Interval		p-value
			Lower	Upper	
(Intercept)	-2.62	1.94	-6.42	1.18	0.18
Output Area Classification Supergroup [Rural residents vs Suburbanites]	2.04	1.31	-0.52	4.60	0.12
Output Area Classification Supergroup [Cosmopolitan vs Suburbanites]	-0.47	1.02	-2.47	1.54	0.65
Output Area Classification Supergroup [Ethnicity central/Constrained city dwellers/Hard-pressed living/Multicultural metropolitans/Urbanites vs Suburbanites]	-0.18	0.88	-1.89	1.54	0.84
Urban/rural classification [Urban conurbation vs Rural]	1.01	1.31	-1.56	3.58	0.44
Urban/rural classification [Urban city and town vs Rural]	0.52	1.04	-1.53	2.56	0.62
Region [East Midlands vs Yorkshire and the Humber]	-2.04	1.57	-5.12	1.04	0.19
Region [East of England vs Yorkshire and the Humber]	0.92	0.98	-1.01	2.85	0.35
Region [London vs Yorkshire and the Humber]	-2.84	1.26	-5.32	-0.37	0.02
Region [North East vs Yorkshire and the Humber]	-0.80	1.71	-4.15	2.55	0.64
Region [North West vs Yorkshire and the Humber]	-1.29	1.21	-3.67	1.09	0.29
Region [South East/South West vs Yorkshire and the Humber]	-2.19	0.96	-4.07	-0.31	0.02
Region [West Midlands vs Yorkshire and the Humber]	-0.30	1.05	-2.36	1.75	0.77
Mixed or single sex [Boys vs Mixed]	-0.29	0.73	-1.72	1.14	0.69
Mixed or single sex [Girld vs Mixed]	0.67	0.58	-0.47	1.80	0.25
Whether school has boarders [Does not have boarders vs Has boarders]	0.24	0.53	-0.79	1.27	0.64
Number of pupils [<701 vs 1,201+]	-0.92	0.73	-2.35	0.51	0.21
Number of pupils [701-1,000 vs 1,201+]	-0.41	0.68	-1.75	0.93	0.55
Number of pupils [1,001-1,200 vs 1,201+]	-0.82	0.81	-2.40	0.76	0.31
Percentage of population with level 4+ qualification in MSOA [Quintile 1 (highest) vs Quintile 5 (lowest)]	2.20	1.08	0.09	4.31	0.04
Percentage of population with level 4+ qualification in MSOA [Quintile 2 vs Quintile 5 (lowest)]	3.44	1.08	1.32	5.56	0.00
Percentage of population with level 4+ qualification in MSOA [Quintile 3 (highest) vs Quintile 5 (lowest)]	2.10	1.00	0.14	4.06	0.04
Percentage of population with level 4+ qualification in MSOA [Quintile 4 (highest) vs Quintile 5 (lowest)]	1.84	1.01	-0.14	3.81	0.07
Percentage of population that own their home in MSOA [Quintile 1 (highest) vs Quintile 5 (lowest)]	-0.74	0.94	-2.59	1.11	0.43
Percentage of population that own their home in MSOA [Quintile 2 vs Quintile 5 (lowest)]	-1.55	0.98	-3.47	0.38	0.11
Percentage of population that own their home in MSOA [Quintile 3 vs Quintile 5 (lowest)]	-1.18	0.92	-2.98	0.63	0.20
Percentage of population that own their home in MSOA [Quintile 4) vs Quintile 5 (lowest)]	1.26	0.77	-0.25	2.77	0.10

Table A.6 Independent schools – young person model Binary logistic regression predicting whether young people sampled from cooperating independent schools participated in the study

Parameter	B	Std. Error	95% Confidence Interval		p-value
			Lower	Upper	
(Intercept)	-3.91	0.96	-5.79	-2.02	0.00
Output Area Classification Supergroup [Rural residents vs Suburbanites]	2.17	0.80	0.60	3.73	0.01
Output Area Classification Supergroup [Cosmopolitan vs Suburbanites]	0.37	0.32	-0.26	1.00	0.25
Output Area Classification Supergroup [Ethnicity central/Constrained city dwellers/Hard-pressed living/Multicultural metropolitans/Urbanites vs Suburbanites]	0.53	0.26	0.02	1.04	0.04
Urban/rural classification [Urban conurbation vs Rural]	1.29	0.79	-0.26	2.83	0.10
Urban/rural classification [Urban city and town vs Rural]	0.37	0.45	-0.51	1.25	0.41
Region [East Midlands vs Yorkshire and the Humber]	-1.41	0.91	-3.21	0.38	0.12
Region [East of England vs Yorkshire and the Humber]	0.59	0.32	-0.05	1.22	0.07
Region [London vs Yorkshire and the Humber]	-1.08	0.72	-2.50	0.33	0.13
Region [North East vs Yorkshire and the Humber]	0.34	1.17	-1.96	2.64	0.77
Region [North West vs Yorkshire and the Humber]	-1.66	0.73	-3.10	-0.23	0.02
Region [South East/South West vs Yorkshire and the Humber]	0.15	0.37	-0.57	0.86	0.69
Region [West Midlands vs Yorkshire and the Humber]	-1.05	0.49	-2.00	-0.10	0.03
Mixed or single sex [Boys vs Mixed]	-0.34	0.30	-0.93	0.25	0.25
Mixed or single sex [Girld vs Mixed]	0.11	0.24	-0.36	0.58	0.65
Whether school has boarders [Does not have boarders vs Has boarders]	0.72	0.21	0.31	1.12	0.00
Number of pupils [<701 vs 1,201+]	0.41	0.26	-0.09	0.92	0.11
Number of pupils [701-1,000 vs 1,201+]	0.79	0.36	0.08	1.51	0.03
Number of pupils [1,001-1,200 vs 1,201+]	-0.30	0.24	-0.77	0.16	0.20
Percentage of population with level 4+ qualification in MSOA [Quintile 1 (highest) vs Quintile 5 (lowest)]	1.59	0.61	0.40	2.78	0.01
Percentage of population with level 4+ qualification in MSOA [Quintile 2 vs Quintile 5 (lowest)]	1.18	0.49	0.21	2.14	0.02
Percentage of population with level 4+ qualification in MSOA [Quintile 3 (highest) vs Quintile 5 (lowest)]	0.97	0.83	-0.65	2.60	0.24
Percentage of population with level 4+ qualification in MSOA [Quintile 4 (highest) vs Quintile 5 (lowest)]	1.75	0.73	0.31	3.19	0.02
Percentage of population that own their home in MSOA [Quintile 1 (highest) vs Quintile 5 (lowest)]	-0.46	0.61	-1.65	0.74	0.45
Percentage of population that own their home in MSOA [Quintile 2 vs Quintile 5 (lowest)]	0.29	0.40	-0.49	1.08	0.46
Percentage of population that own their home in MSOA [Quintile 3 vs Quintile 5 (lowest)]	0.13	0.28	-0.41	0.68	0.63
Percentage of population that own their home in MSOA [Quintile 4) vs Quintile 5 (lowest)]	0.85	0.39	0.08	1.62	0.03

APPENDIX 2 – Weight effectiveness

Table A7. Main study full households (9,330)

	Population	Unwtd (all cases)	Design weighted (all cases)	Final weight (all cases)[18]	Final weight (linked to NPD 6896)[19]
FSM eligibility * SEN status	Percent	Percent	Percent	Percent	Percent
FSM last 6 years & EHC plan	1.9	1.4	1.1	1.9	1.9
FSM last 6 years & other SEND status	4.3	6.4	3.6	4.3	4.4
FSM last 6 years & no SEND status	18.3	34.0	18.8	18.3	18.3
No FSM last 6 years & EHC plan	2.1	1.0	1.4	2.1	2.1
No FSM last 6 years & other SEND status	6.6	4.1	5.9	6.6	6.4
No FSM last 6 years & no SEND status	61.0	51.0	68.9	61.0	61.2
Independent in Y11 and Y12	5.7	2.0	0.3	5.7	5.7
Ethnicity					
Indian	2.7	6.3	3.3	2.7	2.7
Bangladeshi	1.7	6.5	2.2	1.7	1.7
Pakistani	4.2	5.9	4.8	4.2	4.2
Black African	3.8	5.2	3.7	3.8	3.8
Black Caribbean	1.2	3.6	0.9	1.2	1.2
White British / no data	64.9	55.6	69.1	64.9	65
White non-British	5.8	4.4	5.3	5.8	5.6
Mixed / Other	9.9	10.5	10.3	9.9	10
Independent in Y11 and Y12	5.7	2.0	0.3	5.7	5.7
Gender					
Male	48.2	46.8	47.4	48.2	48.1
Female	46	51.1	52.3	46	46.2
Independent in Y11 and Y12	5.7	2.0	0.3	5.7	5.7
Ethnicity * Gender					
Male White British	33.3	26.3	32.5	33.3	33.2
Male Other	15.0	20.5	14.9	15.0	14.9
Female White British	31.7	29.2	36.6	31.7	31.7
Female Other	14.4	21.9	15.6	14.4	14.4
Independent in Y11 and Y12	5.7	2.0	0.3	5.7	5.7

[18] With **W1_MainFamilyFull_weight** applied

[19] With **W1_MainFamily_NPD_weight** applied

	Population	**Unwtd (all cases)**	**Design weighted (all cases)**	**Final weight (all cases)**	**Final weight (linked to NPD (6896))**
KS2 - maths, reading, GPS	Percent	Percent	Percent	Percent	Percent
Upper tertile in all three	14.3	18.3	21.7	14.3	14.3
Upper tertile in two, middle tertile in one	11.8	13.9	15.4	11.8	11.7
Upper tertile in one, middle tertile in two	10.7	11.7	12.3	10.7	10.6
Others with at least one in upper tertile or at least two in middle tertile	24.5	24.7	24.4	24.5	24.2
Lower tertile in two, middle tertile in one	10.2	9.2	8.5	10.2	10.3
Lower tertile in all three	14.1	12.9	10.7	14.1	14.4
Missing data	8.7	7.2	6.7	8.7	8.7
Independent in Y11 and Y12	5.7	2.0	0.3	5.7	5.7
English as an Additional Language					
English is primary language / not recorded	78.4	74.9	83.0	78.4	78.4
English is an additional language	15.9	23.1	16.7	15.9	15.9
Independent in Y11 and Y12	5.7	2.0	0.3	5.7	5.7
School size					
Under 150	22.8	24.8	24.1	22.8	22.6
150-249	53.7	55.8	57.3	53.7	53.9
Over 249	17.8	17.3	18.3	17.8	17.8
Independent in Y11 and Y12	5.7	2.0	0.3	5.7	5.7
School provision					
Special	1.2	0.4	0.7	1.2	1.1
Alternative	0.8	0.7	0.5	0.8	0.8
Selective Other	4.2	5.3	6.1	4.2	4.3
Other	88	91.6	92.4	88	88.1
Independent in Y11 and Y12	5.7	2.0	0.3	5.7	5.7
School region					
East Midlands	8.2	7.9	8.6	8.2	8
East of England	10.6	9.8	11.7	10.6	10.7
London	14.1	19.6	14.4	14.1	13.9
North East	4.4	4.8	4.9	4.4	4.4
North West	13.1	12.8	13.0	13.1	13.1
South East	14.8	13.5	16.5	14.8	14.9
South West	8.8	7.5	9.3	8.8	8.7
West Midlands	10.7	12.5	11.6	10.7	10.8
Yorkshire and the Humber	9.6	9.4	9.7	9.6	9.7
Independent in Y11 and Y12	5.7	2.0	0.3	5.7	5.7

Table A8. Main study Young People (12,828)

	Population	Unwtd (all cases)	Design weighted (all cases)	Final weight (all cases)[20]	Final weight (linked to NPD 9385)[21]
FSM eligibility * SEN status	Percent	Percent	Percent	Percent	Percent
FSM last 6 years & EHC plan	1.9	1.3	1.1	1.9	1.9
FSM last 6 years & other SEND status	4.3	6.3	3.6	4.3	4.3
FSM last 6 years & no SEND status	18.3	33.0	18.8	18.3	18.3
No FSM last 6 years & EHC plan	2.1	1.0	1.3	2.1	2.1
No FSM last 6 years & other SEND status	6.6	4.0	5.8	6.6	6.4
No FSM last 6 years & no SEND status	61.0	49.5	68.6	61.0	61.1
Independent in Y11 and Y12	5.7	4.9	0.8	5.7	5.7
Ethnicity					
Indian	2.7	6.0	3.3	2.7	2.7
Bangladeshi	1.7	5.9	2.0	1.7	1.7
Pakistani	4.2	5.6	4.6	4.2	4.2
Black African	3.8	5.5	4.0	3.8	3.8
Black Caribbean	1.2	3.5	0.9	1.2	1.2
White British / no data	64.9	54.2	68.9	64.9	65.0
White non-British	5.8	4.2	5.3	5.8	5.7
Mixed / Other	9.9	10.2	10.2	9.9	9.9
Independent in Y11 and Y12	5.7	4.9	0.8	5.7	5.7
Gender					
Male	48.2	44.5	46.3	48.2	48.2
Female	46.0	50.6	52.9	46.0	46.1
Independent in Y11 and Y12	5.7	4.9	0.8	5.7	5.7
Ethnicity * Gender					
Male White British	33.3	25.3	31.9	33.3	33.2
Male Other	15.0	19.2	14.4	15.0	14.9
Female White British	31.7	28.9	36.9	31.7	31.7
Female Other	14.4	21.7	16.0	14.4	14.4
Independent in Y11 and Y12	5.7	4.9	0.8	5.7	5.7

[20] With **W1_MainYPFull_weight** applied

[21] With **W1_MainYP_NPD_weight** applied

	Population	Unwtd (all cases)	Design weighted (all cases)	Final weight (all cases)	Final weight (linked to NPD (9385))
KS2 - maths, reading, GPS	Percent	Percent	Percent	Percent	Percent
Upper tertile in all three	14.3	17.6	21.2	14.3	14.3
Upper tertile in two, middle tertile in one	11.8	13.4	15.3	11.8	11.7
Upper tertile in one, middle tertile in two	10.7	11.3	12.2	10.7	10.7
Others with at least one in upper tertile or at least two in middle tertile	24.5	24.2	24.3	24.5	24.5
Lower tertile in two, middle tertile in one	10.2	9.1	8.8	10.2	10.2
Lower tertile in all three	14.1	12.5	10.8	14.1	14.2
Missing data	8.7	7.1	6.7	8.7	8.7
Independent in Y11 and Y12	5.7	4.9	0.8	5.7	5.7
English as an Additional Language					
English is primary language / not recorded	78.4	73.3	82.9	78.4	78.4
English is an additional language	15.9	21.8	16.4	15.9	15.9
Independent in Y11 and Y12	5.7	4.9	0.8	5.7	5.7
School size					
Under 150	22.8	23.8	23.7	22.8	22.7
150-249	53.7	54.4	57.2	53.7	53.8
Over 249	17.8	17.0	18.4	17.8	17.8
Independent in Y11 and Y12	5.7	4.9	0.8	5.7	5.7
School provision					
Special	1.2	0.4	0.7	1.2	1.1
Alternative	0.8	0.7	0.5	0.8	0.8
Selective Other	4.2	5.0	6.0	4.2	4.3
Other	88.0	89.1	92.1	88.0	88.0
Independent in Y11 and Y12	5.7	4.9	0.8	5.7	5.7
School region					
East Midlands	8.2	7.8	8.7	8.2	8.0
East of England	10.6	9.9	11.6	10.6	10.7
London	14.1	19.0	14.5	14.1	13.9
North East	4.4	4.4	4.6	4.4	4.5
North West	13.1	12.3	13.0	13.1	13.1
South East	14.8	13.3	16.5	14.8	14.8
South West	8.8	7.4	9.3	8.8	8.8
West Midlands	10.7	12.2	11.6	10.7	10.7
Yorkshire and the Humber	9.6	8.9	9.5	9.6	9.7
Independent in Y11 and Y12	5.7	4.9	0.8	5.7	5.7