Department
for Education

Labour market outcomes: impact of ethnicity, SES and SEN

Technical report

November 2024

Moira Nelson and Oliver Anderson,
Department for Education

Contents

Introduction

Purpose of technical report

This technical report is a standalone document providing full details on the data used, definitions of variables and methodology used for the Labour market outcomes: impact of ethnicity, socioeconomic status (SES) and special educational needs (SEN) report into the socioeconomic, demographic and education factors affecting labour market outcomes.

The report includes a summary overview and separate chapters on:

- Ethnicity
- Special educational needs
- Socioeconomic status

Each chapter has a separate document and there is a set of accompanying data tables in addition to this technical report.

The Longitudinal Education Outcomes (LEO) dataset

The LEO dataset links information about individuals, including

- personal characteristics such as gender, ethnic group, special educational needs, free school meals eligibility
- education, including schools, colleges and higher education institutions attended, courses taken and qualifications achieved
- employment and income
- benefits claimed

More detail on the source datasets, data matching processes and data quality for the LEO data used in this report can be found in the technical report[1] accompanying Post-16 education and labour market activities, pathways and outcomes (LEO) - GOV.UK.

[1] Technical Report for Education and Labour Market Pathways of Individuals (LEO)

Data definitions

Labour market outcome variables

Two labour market outcome variables have been defined for this analysis, which are designed to look at outcomes from an economic sense i.e. the value to the Exchequer.

Labour market outcomes are measured in the 2017-18 tax year. As multiple cohorts of school leavers are examined in this analysis (see Definition of cohorts) this means that labour market outcomes are measured between 8 and 15 full tax years after key stage 4 (ages 24 to 31), depending on the cohort.

Definition of good outcome

A good labour market outcome is defined (for the purposes of this analysis) as being in paid employment (PAYE) for at least one day of each month of the tax year, and an upper quartile earner in this year.

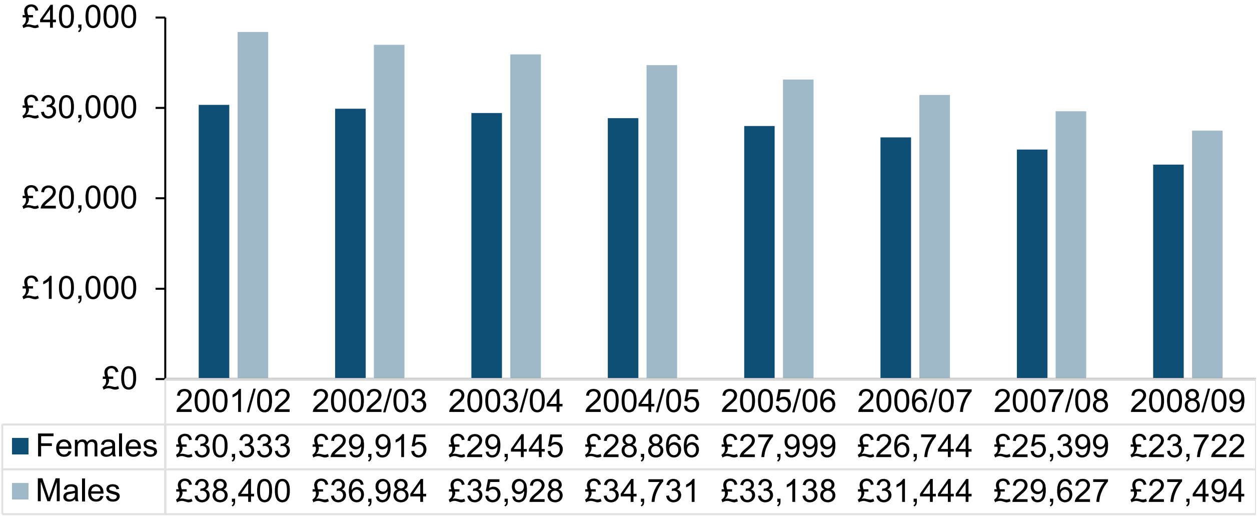
Earnings from self-employment are not included in the definition of good outcome used in the main models shown in this report.

Earnings used for the upper quartile calculations are annualised: earnings reported in the tax year are divided by the number of days recorded in the employment spells in the tax year for that individual. This provides an average daily wage, which is then multiplied by the number of days in the tax year to create their annualised earnings[2].

Earnings quartiles are calculated for **all individuals** in the cohorts examined, and separately for each cohort and by gender. See Table 1 for the thresholds used to identify upper quartile earners. Earnings for an individual in the 2017-18 tax year are compared to the upper quartile threshold. LEO PAYE data does not include an indicator of hours worked by an individual which means we are unable to identify those who work part-time and therefore cannot adjust for this in this analysis. This may account for some of the difference in upper quartile thresholds between males and females. Around 15 per cent of males and females meet this definition.

[2] Note: we do not know the actual number of hours worked just the length of the employment spells, so this method does not adjust for the number of hours worked per day, or the number of days worked per week. For example, if an individual is employed for the full tax year we will use 365 days (or 366 in leap year) in the calculation.

Table 1: Upper quartile earnings thresholds in 2017-18 tax year: females and males by KS4 cohort



	2001/02	2002/03	2003/04	2004/05	2005/06	2006/07	2007/08	2008/09
■ Females	£30,333	£29,915	£29,445	£28,866	£27,999	£26,744	£25,399	£23,722
■ Males	£38,400	£36,984	£35,928	£34,731	£33,138	£31,444	£29,627	£27,494

Source: Authors' analysis using Longitudinal Education Outcomes data

Alternative definitions of good outcome

Alternative definitions for the good outcome measure have been calculated as robustness checks on the definition as follows:

1. Additional analysis using **total PAYE and self-employment income** was conducted for comparative purposes. This uses PAYE income from all individuals with any employment spells in the tax year (rather than one day in each month). Where an individual has income from both employment spells paid through PAYE and through self-employment, the earnings used is the sum of their PAYE earnings and their total earnings from self-employment. Upper quartile thresholds were calculated separately for each cohort and for males and females. Around 20 per cent of males and females meet this definition.

Alternative definitions measuring good labour market outcomes at the same age for multiple cohorts (rather than in a single tax year) were also conducted. Annualised earnings from PAYE are included for those with employment in each month of the relevant tax year and upper quartile thresholds were calculated separately for each cohort and for males and females.

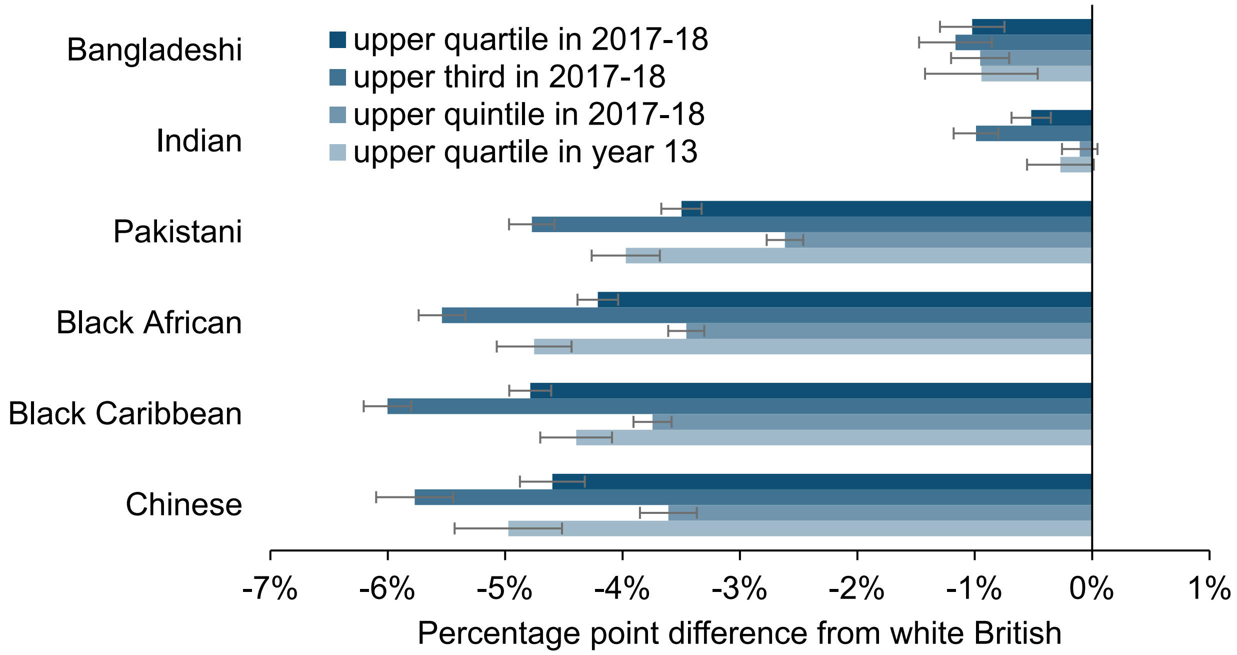
2. **Upper quartile earners 8 full tax years** after key stage 4 (approximate age 24) – this was calculated for all 8 cohorts. Around 15 per cent of males and females meet this definition.
3. **Upper quartile earners 13 full tax years** after key stage 4 (approximate age 29) – this was calculated for the first 3 cohorts only. Around 15 per cent of males and females meet this definition.

Alternative definitions for the good outcome measure using different earnings thresholds for the 8 cohorts were also derived. As in the main measure, annualised earnings from PAYE are included for those with employment in each month of the 2017-18 tax year and upper quartile thresholds were calculated separately for each cohort and for males and females.

4. **Upper third earners in 2017-18 tax year**. Earnings tertiles were calculated for all individuals in the 8 cohorts. Around 20 per cent of males and females meet this definition.
5. **Upper quintile earnings in 2017-18 tax year**. Earnings quintiles were calculated for all individuals in the 8 cohorts. Around 12 per cent of males and females meet this definition.

Probit regressions using these alternative definitions were carried out. Average marginal effects for ethnicity after adding all socioeconomic, demographic and education controls are shown for a selection of these in Figure 1.

Figure 1: Males – marginal effects of ethnic group on alternative definitions of good labour market outcome



Error bars represent 95% confidence intervals

Source: Authors' analysis using Longitudinal Education Outcomes data

It can be seen that although the gaps between ethnic groups vary slightly for each measure, the trends are very similar to those seen with the definition of good outcome used in the report (upper quartile in 2017-18) measure. Similar trends to the chosen good outcome measure were also seen for special educational needs and socioeconomic status.

Definition of poor outcome

A poor labour market outcome is defined as claiming out-of-work benefits for at least one day in each of (at least) 6 consecutive months of the tax year (2017-18). Around 8 per cent of males and 12 per cent of females meet this definition. The benefits classed as out-of-work for this analysis were:

- Jobseekers Allowance (JSA)
- Jobseekers Training Allowance (JTA)
- Employment and Support Allowance (ESA)
- Incapacity Benefit (IB)
- Income Support (IS)
- Passported IB (PIB)
- Severe Disablement Allowance (SDA)
- Pension Credit (PC)
- State (Retirement) Pension (RP)
- Carers Allowance (Invalid Carers Allowance – ICA)
- Attendance Allowance (AA)
- Universal Credit – Searching for Work (UAA)
- Universal Credit – No Work requirements (UBC)
- Universal Credit – Preparing for work (UCE)
- Universal Credit – Planning for work (UDF)

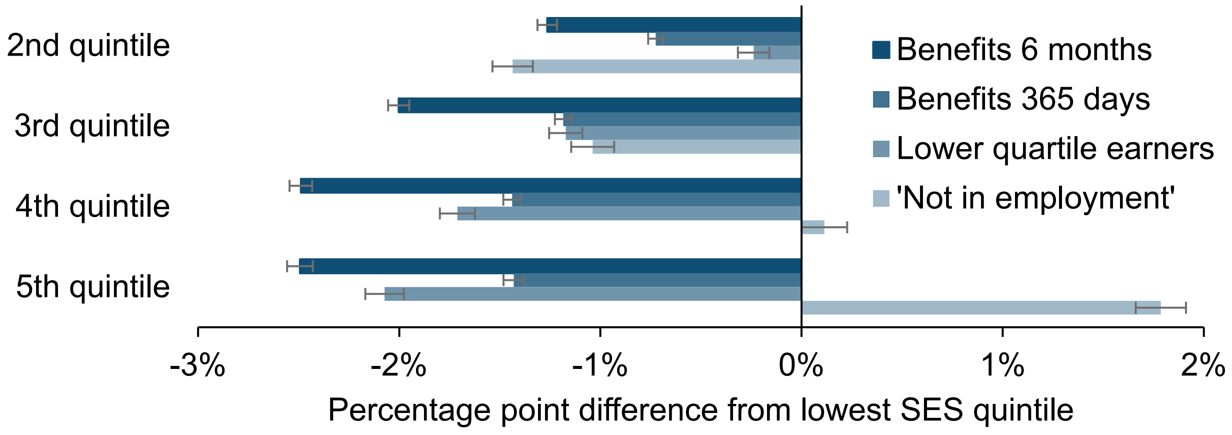
Alternative definitions of poor outcome

Alternative definitions for the poor outcome measure have been calculated as robustness checks on the definition as follows:

1. Outcome measured **8 full tax years** after key stage 4 (approximate age 24) – this was calculated for all 8 cohorts. As per the main measure, the outcome is defined as one day in each of (at least) six months. The outcome tax year is different for each cohort. Around 9 per cent of males and 13 per cent of females meet this definition.

2. Outcome measured **13 full tax years** after key stage 4 (approximate age 29) – this was calculated for the oldest 3 cohorts. As per the main measure, the outcome is defined as one day in each of (at least) six months. The outcome tax year is different for each cohort. Around 7 per cent of males and 12 per cent of females meet this definition.
3. Out-of-work benefits claims cover **365 days of the 2017-18 tax year**. This was calculated for all individuals in the cohorts examined. Around 5 per cent of males and 8 per cent of females meet this definition.
4. Out-of-work benefits claims cover a continuous period of at least **90 days in the 2017-18 tax year**. This was calculated for all individuals in the cohorts. Around 9 per cent of males and 13 per cent of females meet this definition and includes everyone in the main poor outcome definition.
5. **Lower quartile earners in 2017-18 tax year**. Earnings quartiles were calculated for all individuals in the cohorts examined, and separately for each cohort and by gender. Earnings from PAYE are included for those with employment in each month of the tax year. Around 15 per cent of males and females meet this definition.
6. Those who **do not meet the one day in each month criterion for employment**, regardless of earnings. This will include those not in steady employment as well as those not actively in the labour market such as students, financially independent individuals, stay at home parents and those are actively seeking work. Over a third of males and females meet this definition.

Figure 2: Males – marginal effects of socioeconomic status on alternative definitions of poor labour market outcome



Error bars represent 95% confidence intervals

Source: Authors' analysis using Longitudinal Education Outcomes data

Probit regressions using these alternative definitions were carried out. Average marginal effects for socioeconomic status, after adding all demographic and education controls, are shown for a selection of these in Figure 2.

The gaps between SES quintiles vary slightly between definitions which measure benefits spells but the trends are very similar. The trend for 'lower quartile earners' (definition 5) is less similar, but those who meet this definition are in regular employment and therefore could be thought of as in a better labour market outcome than those on benefits for a sustained period. The trends for 'not in employment' (definition 6) are quite different, with those from the least disadvantaged background having a higher chance to meet this definition that those from the most disadvantaged, suggesting that this definition is too heterogenous for this purpose. Similar trends to the selected poor outcome measure were seen also for special educational needs and ethnicity.

Explanatory variables and controls

The following variables are used as explanatory variables and controls, as breakdowns for descriptive statistics, or used in the calculation of the dependent variable.

Gender – Used in the calculation of income quartiles for good outcome. This is taken from key stage 4 National Pupil Database (NPD) data and is available for all individuals regardless of school type. For individuals in school types without a pupil level census (e.g. independent schools) this is taken from data provided by awarding organisations.

Ethnicity in year 11 – Minor ethnic group taken from NPD data for the academic year the individual finished key stage 4. This is not available for those individuals in school types without pupil level data collections (e.g. independent schools). In the 2001/02 academic year, schools were using two different classification systems. The majority of schools were using an older classification which did not have any mixed ethnicities. Instead, these were classed as Any Other Ethnic Group. Furthermore, there was no breakdown for Traveller of Irish Heritage or Gypsy/Roma in 2001/02 (classed as other White). Due to small numbers, the codes for Irish Traveller and Gypsy/Roma have been combined into a single group. See also the Backfilling and imputation of characteristics methodology used to increase coverage.

Socioeconomic status (SES) quintile in year 11 – quintiles of socioeconomic status were created from an index of socioeconomic status combining an individual's free school meals (FSM) status, the Index of Multiple Deprivation[3] score and three local area-based measures (proportion of individuals in higher or lower managerial or professional occupations, proportion of individuals whose higher level of achievement is level 3 or above, proportion of individuals who own their home). The bottom quintile (1) is those

[3] English indices of deprivation - GOV.UK (www.gov.uk)

with the lowest socioeconomic status (most deprived) and the top quintile (5) is those with the highest socioeconomic status (least deprived). This is calculated for all individuals regardless of school type, though we find that those individuals who attended independent schools are primarily in the top SES quintile. Therefore the top quintile is slightly smaller for the state-funded school population. It is also common to use FSM eligibility or the Income Deprivation Affecting Children Index (IDACI) but there are issues with using either of those and this report uses the derived SES measure based on that used by the Institute for Fiscal Studies[4].

Special Educational Needs (SEN) in year 11 – This is taken from NPD data for the academic year the individual finished key stage 4. SEN analysis is broken down into three sub-groups: 1) individuals with statement of SEN, 2) individuals with SEN without a statement and 3) individuals not identified as SEN. This uses the SEN Code of Practice[5] which came into effect on 1 January 2002, before the introduction of Education, Health and Care (EHC) plans. Under this Code of Practice, a child or young person could be identified in one of three categories: statement of SEN, School Action or School Action Plus. A statement of SEN is when a formal assessment has been made which sets out the child's need and the extra help they should receive. For this analysis, the SEN categories School Action and School Action Plus are combined into 'SEN without statement'. See also the Backfilling and imputation of characteristics methodology used to increase coverage.

English as an additional language (EAL) in year 11 - EAL is primarily taken from NPD data for the academic year the individual finished key stage 4. Again, as with other variables stated above, this is not available for those individuals in school types without pupil level data collections (e.g. independent schools). See also the Backfilling and imputation of characteristics methodology used to increase coverage.

Local authority of residence in year 11 – local authority in year 11/age 16 is derived small area statistics information from either School Census or from geographical information available in the LEO dataset.

Local authority of residence in 2017-18 – local authority is derived from small area geographic information in the LEO dataset for the 2017-18 tax year. The new local authority is recorded if this is different from the local authority of residence in year 11 (otherwise 'Did not move').

[4] Widening participation in higher education: analysis using linked administrative data - Chowdry - 2013 - Journal of the Royal Statistical Society: Series A (Statistics in Society) - Wiley Online Library

[5] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/273877/special_educational_needs_code_of_practice.pdf

Travel to work areas[6] (TTWA) in year 11 and 2017-18 – these were derived using similar methodology to the two local authority variables. Probit regression analysis using TTWA rather than local authority was carried out but this had little effect on the results.

Region of school – this is derived from the administrative local authority for the school the individual attended when they were in key stage 4 and is available for all individuals. This is the region of the school and may not be where the pupil lived (including pupils residing in Scotland or Wales who cross borders to attend school).

Season of birth – this is derived from the month of birth available in the KS4 NPD data. There are three seasons: Autumn (September to December), Spring (January to April), Summer (May to August)

Key stage 2 (KS2) attainment – the level of achievement in key stage 2 statutory assessments in maths and in English carried out in English state schools at age 11. A dummy is used to denote where KS2 data in not available for an individual.

Key stage 4 attainment – this is taken from the key stage 4 NPD data and is therefore available for all individuals. The measure used is the total performance points for all GCSEs and equivalent qualifications. The equivalents which counted in each academic year, and therefore for each cohort, may differ depending on performance tables rules at the time.

Highest level of education – the highest level[7] of education achieved by the 2017/18 academic year in English schools, English further education institutions and UK higher education institutions. These are used in the report as follows:

Below level 2 – did not achieve qualifications equivalent to full level 2
Level 2 – highest level of qualification was equivalent to full level 2 (equivalent to 5 GCSEs)
Level 3 – highest level of qualification was equivalent to full level 3 (equivalent to 2 A levels)
Levels 4 to 8 – any qualification at this level

School type - the school type for each individual is taken from the key stage 4 NPD data. This is therefore available for all individuals. Many school types are newer than the oldest cohorts in this data, for example sponsored academies and free schools.

School progress score – the value-added measures for schools calculated for school performance[8] are not consistent across the academic years in which the cohorts in this analysis did their GCSEs. A KS2 to KS4 progress score has been calculated which is

[6] Travel to work area analysis in Great Britain - Office for National Statistics
[7] What qualification levels mean: England, Wales and Northern Ireland - GOV.UK (www.gov.uk)
[8] Compare the performance of schools and colleges in England - GOV.UK

based on the principle of Progress 8[9] but which uses the KS2 and KS4 measures of the time and is consistent across the cohorts of interest.

School demographics – the percentage of individuals in each cohort in each school eligible for free school meals, with a statement of SEN, with SEN without a statement and the urban/rural status of the school. This also includes the percentage of individuals in each cohort in each school who achieved at least 5 A*-C at GCSE. Some of these variables are only available for school types with pupil level data collections.

Higher education variables – these are derived from HESA data for those who attended higher education in the UK where higher education is the highest level of education. Data items include the classification of first degree, the subject group of the highest level of degree and the institution type attended for the highest level of degree awarded. Five subject groups are derived from earnings returns for each subject by degree type and gender[10][11] split into quintiles.

Subject area at further education – subject sector area (SSA) for the latest achievement at the highest level is taken from the ILR data for those whose highest level of education is through either an apprenticeship or classroom learning in the English further education system. Five subject groups are derived from average earnings returns for each SSA by gender[12] split into quintiles.

A levels at 16-18 variables – these are derived from key stage 5 (KS5) NPD data and include total performance points for A levels and other academic qualifications, and A level subjects. Five A level subject groups for males and females are derived from median earnings at age 27 for individuals with a pass in each subject, split into quintiles. Students generally study three A level subjects: we do not distinguish between these for earnings potential at an individual level – rather we look across the cohort to analyse the impact of A level choice on average earnings. This implicitly assumes that individuals take subject combinations with similar earnings potential, which will not necessarily be true in all cases.

Backfilling and imputation of characteristics

The main demographic characteristics examined in this report have initially been taken from the value assigned in the pupil level census data for the academic year the individual completed key stage 4. Some of these data items are missing or incomplete, or are not available for individuals attending certain school types. A backfill and imputation

[9] Secondary accountability measures (including Progress 8 and Attainment 8) - GOV.UK (www.gov.uk)
[10] Undergraduate degrees: labour market returns - GOV.UK
[11] Postgraduate degrees: labour market returns - GOV.UK
[12] Labour market outcomes disaggregated by subject area using the Longitudinal Education Outcomes (LEO) data

methodology has been used to complete some of this information using NPD data from other academic years, or ILR and HESA data where this is available and appropriate.

EAL – if information on EAL is missing for year 11, this is backfilled using school census data from other years, taking the value for the year closest to the key stage 4 year. If this data item is still missing, the value is imputed using ethnicity.

Ethnicity – if information on minor ethnic group is missing for year 11, the value for ethnicity from the academic year closest to the key stage 4 year from school census, ILR or HESA is used. Where multiple sources have a value for the same (closest) year, the school census value is taken if present, then ILR and finally HESA. Note that HESA data does not contain a code for white British. Code 10 (white) has not been used in the backfilling process. No imputation has been used for ethnicity: instead, a 'not known' category has been included where all individuals are included.

SEN backfilled – if information on special educational needs is missing for year 11, this is backfilled using school census data for other years, taking the value for the closest year to the key stage 4 year. Where there are multiple values for SEN status for the closest year to KS4, a hierarchy is used (SEN with a statement, SEN without a statement, not identified with SEN). Backfilling from ILR and HESA has not been carried out as the data items are too dissimilar. Imputation has not been used.

Table 2: Backfilling and imputation of EAL, ethnicity and SEN – percentage of individuals with missing data items

	EAL All schools	EAL State-funded schools	Ethnicity All schools	Ethnicity State-funded schools	SEN All schools	SEN State-funded schools
School census yr 11	8.6%	0.7%	10.9%	3.3%	8.4%	0.5%
After backfilling	5.5%	0.1%	2.9%	0.1%	5.5%	0.1%
After imputation	0.3%	0.0%	n/a	n/a	n/a	n/a

Source: Authors' analysis using Longitudinal Education Outcomes data

The percentage of individuals with missing data items at each stage of the backfill and imputation methodology is shown in Table 2 for each variable. The main model in the report (for individuals who attended state-funded schools) uses the backfilled versions of these data items. The imputed versions have been used for the robustness check with pupils from all school types (see Alternative model in the Methodology section).

Methodology

Definition of cohorts

Individuals are assigned to the latest academic year they appear in the key stage 4 data. This academic year is referred to as the 'cohort'. Most individuals will be academic age[13] 15, but a minority are older or younger than this depending on their circumstances.

The earliest cohort in the dataset finished key stage 4 in the 2001/02 academic year. This is the earliest pupil-level data available in the NPD. The employment, earnings and benefits data are examined in the 2017-18 tax year, which means that the labour market outcomes for the 2001/02 cohort are measured 15 full tax years after key stage 4, and 8 full tax years for the 2008/09 cohort.

Cohorts of individuals who completed key stage 4 in England between 2001/02 and 2008/09 have been combined to produce a more representative and robust picture of individuals' education and labour market outcomes. When looking at all eight cohorts of individuals, this consists of 4.9 million individuals across all school types matched to the LEO data (4.5 million in state-funded schools). This is particularly important when looking at smaller sub-groups of interest. Combining cohorts of individuals completing their GCSEs at the same time means any changes or patterns are more likely to be real differences (and not say, something randomly different about a certain group from a certain year).

Coverage

The analysis in the main model has only been carried out for those who did their GCSEs in a state-funded school in England. These are school types in which there was a pupil level school census (in order to obtain the pupil level characteristic data) in the academic years covered by the analysis. These are:

- Local authority maintained mainstream schools
- Local authority maintained special schools
- Sponsor-led academies
- City technology colleges

Those whose school type at the end of KS4 differs from the above have been removed from the analysis for the main models.

[13] Age at the start of the academic year i.e. 1st September

Those with missing individual data items after backfilling (such as ethnicity or SES quintile) have also been removed for the main analysis. Those with missing school data items (e.g. school demographics) have been removed for the ethnicity and socioeconomic status analyses. See Table 3 for a breakdown of the number of individuals by KS4 cohort.

Table 3: Number of individuals in analytical population

KS4 cohort	Number of individuals	Number of individuals matched to LEO data	Number in state-funded schools	Number with no missing individual data items (note 1)	Number with no missing school information (note 2)
2001/02	589,516	547,747	505,956	501,338	499,543
2002/03	621,676	580,383	539,474	537,408	533,030
2003/04	637,169	602,492	558,853	557,707	553,684
2004/05	641,516	610,239	559,866	558,541	554,696
2005/06	654,730	627,467	574,722	573,743	569,869
2006/07	660,746	637,152	582,771	582,103	578,318
2007/08	659,066	639,560	584,474	583,728	579,952
2008/09	642,076	626,050	570,064	569,423	565,601
Total	5,106,495	4,871,090	4,476,180	4,463,991	4,434,693

Notes:

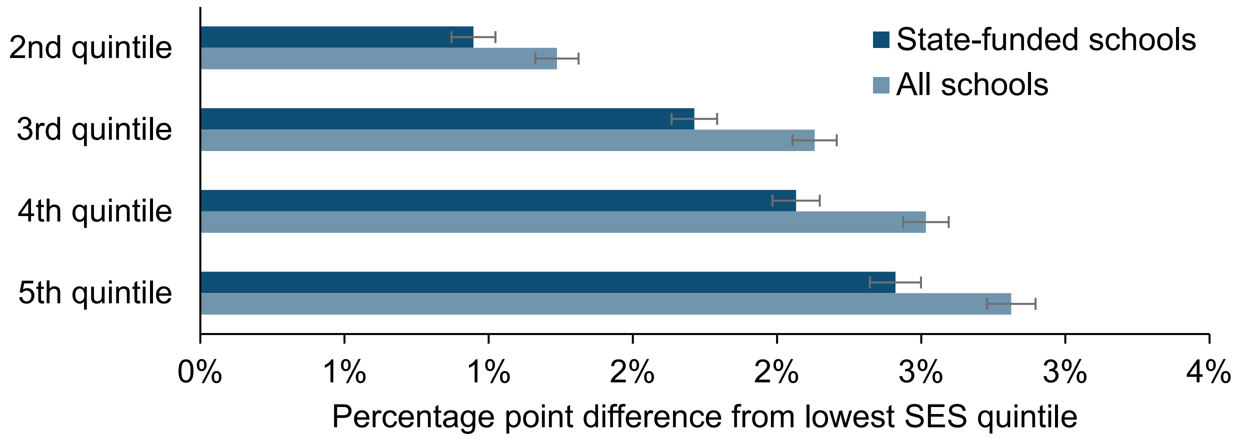
1. These individuals were included in the analysis in the SEN chapter
2. These individuals were included in the analysis in the ethnicity and SES chapters

Source: Authors' analysis using Longitudinal Education Outcomes data

Alternative model

Individuals from state-funded schools were used in the analysis in order to make use of the pupil level data available. However, the probit regression analysis was also carried out on individuals in the 8 GCSE cohorts from all school types, using the backfilled and imputed versions of the pupil characteristics data (see Figure 3 for a comparison of the average marginals effects for socioeconomic status, after adding all demographic and education controls to the main model). Where data items are still missing, a 'missing' category has been added. This model does not contain any school level controls as these were not available for all school types.

Figure 3: Males – marginal effects of socioeconomic on good labour market outcome with different school types



Error bars represent 95% confidence intervals

The trends between SES quintiles for those from all schools are very similar to those from state-funded schools (main model), suggesting that the omission of those who attended schools which did not have a pupil level collection (such as independent schools) does not change the findings of the analysis. Similar trends to the state-funded model were also seen for ethnicity and special educational needs.

Descriptive analysis

A: Proportions by ethnicity, SES quintile and SEN group

The percentage of all individuals (in the 8 state-funded cohorts) by ethnic group, SES quintile and SEN status was calculated. Percentages are shown for males and females separately and in total. The results can be found in tables 1.1[14], 2.1 and 3.1 in the accompanying tables for the appropriate chapter. Results are shown for all ethnic groups.

In addition, breakdowns showing the composition of each SES quintile with regard to a number of variables can also be found in table 2.x. This corresponds to the charts shown and discussed in Section 2A of the SES chapter.

[14] In the accompanying tables, table names are shown with an underscore rather than decimal point (i.e. Table_1_1 rather than Table 1.1

B: Observed labour market outcomes

The percentage of individuals in each ethnic group, SES quintile and SEN status group in **good outcome** has been calculated. Further breakdowns by region of school, highest level of education, ethnicity, socioeconomic status, special educational needs, degree classification and higher education (HE) institution type are also provided, broken down by gender and for all individuals. The results can be found in tables 1.2, 2.2 and 3.2 in the accompanying tables for the appropriate chapter. All ethnic groups are included in the ethnicity tables.

Equivalent tables are provided for **poor outcome** and can be found in tables 1.3, 2.3 and 3.3 in the accompanying tables for the appropriate chapter.

These tables correspond to the charts shown in Section 2 of the ethnicity and SEN chapters and Section 2B of the SES chapter.

Regression analysis

As the outcome variables being measured in this analysis can only take two values (either in a good/poor outcome or not), a binary regression model was used to estimate the probability that an individual will fall into one of these categories.

Probit regression is one such binary regression methodology. This predicts the probability of a non-zero response e.g. the probability of an individual falling into the good outcome category, or where Y=1 in the equation below. The probit methodology models the inverse standard normal distribution of this probability as a linear combination of the predictor variables.

Equation 1: Probit regression

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_k \mathrm{X}_k + \mathrm{u}$$

with

$$\mathrm{P}(\mathrm{Y} = 1 | \mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_k) = \Phi(\beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_k \mathrm{X}_k)$$

Equation 1 shows the population Probit model with multiple regressors X_1 to X_k where Y is the binary outcome variable, P is the probability and Φ is the cumulative standard normal distribution function.

A probit model was selected over a logit model (which uses a logistic distribution rather than a normal distribution). There is little difference in practice between the two methodologies. As a check, regression analysis using a logit model rather than probit was carried out with ethnic group as the dependent variable. This produced results with very similar trends.

The coefficients produced using a probit regression represent, for continuous variables, the change in z score for a one-unit change in the predictor variable. For a categorical predictor variable (such as ethnicity, SES quintile and SEN status used in this report) this represents the change in z score between the variable and the reference value e.g. between ‘statement of SEN’ compared to ‘no SEN’.

Average marginal effects (AME) have been calculated from the probit coefficients. This represents the discrete difference in probability between the variable and the reference value. For average marginal effects, these are calculated for each observation in the dataset and then averaged. It represents the average difference in the probability of being in the outcome group associated with a one-unit change in the variable of interest relative to its reference category.

For each of the three independent/explanatory categorical variables (ethnic group, SES and SEN), probit regression was performed with both good outcome and poor outcome as the outcome variable for males and females separately (Equation 2). This was performed in STATA using *probit*, followed by *margins* command with *dydx* option to give the AME.

Equation 2: Analytical probit regression model

$$\mathrm{P(Outcome\_Var} = 1) = \Phi(\beta_0 + \beta_1 \mathrm{Ind\_Var} + \beta_2 \mathrm{Demographics} + \beta_3 \mathrm{Education})$$

Where:

Outcome_Var: the outcome variable, either good outcome or poor outcome

P is the probability that Outcome_Var =1

Ind_var: the independent variable, either ethnic group, SES or SEN

Ethnicity– categorical variable with 17 values. Reference: white British
SES – categorical variables with 5 values. Reference: 1st quintile
SEN – categorical variable with 3 values. Reference: no SEN

- Demographics: a vector of socioeconomic and demographic controls:
    - Ethnicity (not for ethnicity model)
    - SES (not for SES model)
    - SEN (not for SEN model)
    - EAL
    - Current LA
    - LA during school
    - Season of birth
    - KS4 cohort

- Education: a vector of education controls
    - School demographics (not for SEN model):
        - School type
        - School progress score
        - Percentage of pupils eligible for FSM in KS4 school
        - Percentage of pupils with a statement of SEN in KS4 school
        - Percentage of pupils with SEN without a statement in KS4 school
        - Percentage of pupils in school cohort achieving 5 A*-C
        - Urban/rural indicator
    - Pre-16 education
        - KS2 maths level
        - KS2 English level
        - Total performance points at KS4
    - Post-16 education
        - Highest level of education
        - A level subjects studied
        - KS5 academic points
        - FE classroom subject
        - Apprenticeship subject
        - Degree classification
        - Subject at higher education
        - HE institution type

These results are discussed in the 'Effect of introducing controls' section of the relevant chapter. Full results for the probit regression models are also included in the accompanying tables for the appropriate chapter (Tables 1.4 to 1.7 for ethnicity, tables 2.4 to 2.7 for socioeconomic status and tables 3.4 to 3.7 for special educational needs).

Table 4: Example row from table of probit regression results

Ethnicity	**State-funded population** **Good Outcome** **No controls**	**State-funded population** **Good Outcome** **All controls**
Bangladeshi	-0.030*** (0.002)	-0.020*** (0.003)

Source: Authors' analysis using Longitudinal Education Outcomes data

An example row of regression results in shown in Table 4. The first column shows the group being compared to the reference group (in this case, the reference is white British). The second column shows the difference in observed probability of good outcome between the Bangladeshi group and white British (a difference in probability of -0.030, or 3 percentage points lower) and that this difference is highly significant as indicated by the three asterisks. The figure in brackets is the standard error. The third column shows the difference in probability of good outcome between the Bangladeshi group and white British after controlling for all socioeconomic, demographic and education variables (a difference in probability of -0.020, or 2 percentage points lower). Again, we can see that this is highly significant.

Statistical significance is denoted with asterisks (*** denotes P <0.001, ** denotes P<0.01, * denotes P<0.05).

Decomposition analysis

Decomposition analysis is a multivariate technique which is used to understand the average difference in outcomes between two groups by breaking these down into components and quantifying the contribution of individual factors to this difference. The first component is the part attributable to compositional differences in groups (e.g. education levels). The second is the part attributable to variations in the effects of those characteristics (e.g. differences in the returns or behavioural responses for those with the same characteristics).

Linear multivariate decomposition techniques are commonly known as Oaxaca-Blinder models. However, as the outcome variable in this analysis is binary rather than linear, the decomposition is used with a probit regression model.

Decomposition methodologies use regression to measure the difference in predicted outcomes of between a comparator high outcome group (A) and a reference low outcome group (B). The dependent outcome variable is a function of a linear combination of predictors and regression coefficients. In this analysis, probit regression is used so Y can be expressed as in Equation 3 (a simplified version of Equation 1).

Equation 3: Simplified probit regression

$$\mathrm{Y} = \Phi(\mathrm{X}\beta)$$

Where Y is the dependent variable, X is the matrix of explanatory variables, β is the vector of coefficients. Using this, the mean difference in the outcome variable (Y) between groups A and B can be decomposed as per Equation 4:

Equation 4: Decomposition of the mean difference

$$\overline{\mathrm{Y}}_{\mathrm{A}} - \overline{\mathrm{Y}}_{\mathrm{B}} = \overline{\Phi(\mathrm{X}_{\mathrm{A}}\beta_{\mathrm{A}})} - \overline{\Phi(\mathrm{X}_{\mathrm{B}}\beta_{\mathrm{B}})}$$

$$= \underbrace{\left\{\overline{\Phi(\mathrm{X}_{\mathrm{A}}\beta_{\mathrm{A}})} - \overline{\Phi(\mathrm{X}_{\mathrm{B}}\beta_{\mathrm{B}})}\right\}}_{\mathrm{E}} + \underbrace{\left\{\overline{\Phi(\mathrm{X}_{\mathrm{A}}\beta_{\mathrm{A}})} - \overline{\Phi(\mathrm{X}_{\mathrm{B}}\beta_{\mathrm{B}})}\right\}}_{\mathrm{C}}$$

Where A is the high outcome group and B is the reference group and:

E – **Characteristics** component: this is sometimes referred to as Endowments and is the counterfactual comparison of the difference in outcomes from group A's perspective i.e. this is the expected difference if A were given B's distribution of characteristics

C – **Returns** component: this is sometimes referred to as Coefficients and is the counterfactual comparison of differences in outcomes from group B's perspective i.e. this is the expected difference if B experienced A's behavioural responses

These components are then further decomposed into the mean difference in predicted outcomes attributable to each explanatory variable (or factor), and a constant term (referred to as **Unexplained** in this report) representing that part of the difference which is not explained by the explanatory variables.

The decomposition was carried out in STATA with the *mvdcmp*[15] command using probit as the estimation method using Equation 5, with a grouping variable used to identify the high and low outcomes groups.

Equation 5: Probit regression model used for decomposition using socioeconomic status and good outcome as an example

$$\overline{\mathrm{Y}}_{\mathrm{SES}_5} - \overline{\mathrm{Y}}_{\mathrm{SES}_1} = \overline{\Phi(\mathrm{X}_{\mathrm{SES}_5}\beta_{\mathrm{SES}_5})} - \overline{\Phi(\mathrm{X}_{\mathrm{SES}_1}\beta_{\mathrm{SES}_1})}$$

where Y is the probability of good outcome and

[15] Mvdcmp: Multivariate Decomposition for Nonlinear Response Models

$$\begin{aligned}\Phi(X\beta) = \Phi(&\beta_0 + \beta_1\text{Ethnicity} + \beta_2\text{SEN} + \beta_3\text{EAL} + \beta_4\text{KS4Region} + \beta_5\text{KS2Maths}\\ &+ \beta_6\text{KS4TotPts} + \beta_7\text{SchoolType} + \beta_8\text{SchoolProgress} + \beta_9\text{SchoolDemographics}\\ &+ \beta_{10}\text{Post16Education})\end{aligned}$$

For the decomposition of ethnicity outcome comparisons, SES is used in place of Ethnicity in Equation 5. For SEN comparisons, SES in used in Equation 5 in place of SEN and there are no school explanatory variables.

The decomposition has been carried out for the pairs of groups shown in Table 5 for good outcome and for poor outcome, for both males and females separately. In each case, the group with a higher chance of good outcome (or a higher chance of poor outcome) is treated as the comparator group and the group with the lower chance of good (or poor) outcome is the reference group.

Table 5: Decomposition pairs

Group	Reference	Comparator
Ethnicity	white British	Bangladeshi
Ethnicity	white British	Indian
Ethnicity	white British	Pakistani
Ethnicity	white British	black African
Ethnicity	white British	black Caribbean
Ethnicity	white British	Chinese
Socioeconomic status	1st quintile	5th quintile
Special educational needs	No SEN	SEN without a statement
Special educational needs	No SEN	Statement of SEN

The contribution of each explanatory variable (or factor) within each component (Characteristics or Returns) can be expressed as a percentage of the whole difference in mean outcomes between the two groups. Thus, the contributions are additive, so for each of interpretation, some related factors have been grouped together.

The factors or factor groups presented are:

- Ethnicity (not for ethnicity decomposition)
- SES (not for socioeconomic status decomposition)
- SEN (not for special educational needs decomposition)

- EAL
- Region at KS4
- KS4 school (school type, school demographics and school progress – not included in SEN decomposition)
- Pre-16 attainment (KS2 maths level, KS4 total performance points)
- Post16 education[16] – categories grouped as follows:
    - SES and good outcome for ethnicity – below degree, lower second class or below, first or upper second class, postgraduate
    - Poor outcome for ethnicity – below degree, post-92, pre-92, Russell group, other HE
    - SEN – level 2, level 3, level 4/5, level 6+

These results are discussed in the 'Relative importance of controls' sections of the relevant chapters and are presented in table form in the tables accompanying each chapter (Tables 1.8 to 1.11 for the ethnicity chapter, Tables 2.8 to 2.11 for SES and Tables 3.8 to 3.11 for SEN). Within the text of each chapter, the interpretation of the charts and tables is fully explained.

The probit models used for the decomposition are based on the models used for the main probit regressions, but there are important differences. The probit regressions have a single explanatory variable (ethnicity, SES or SEN) and a vector of controls; in the decomposition all variables are explanatory variables. This makes the analysis more susceptible to collinearity so some variables (e.g. KS2 English level) have been omitted to avoid this. This may lead to differences in the explanatory power of the decomposition compared to the probit regression for the same comparison.

In addition, some of the groups are relatively small, such as the Chinese ethnic group and the statement of SEN group. These groups do not necessarily have observations for some values of the possible explanatory variables which would lead to omitted observations in a probit model.

There are also stark differences in the distributions of characteristics for some groups. This is particularly true for some minority ethnic groups when compared to the majority white British group (around 85 per cent of the individuals in the analysis). This means that the white British group generally has a good distribution of each explanatory variable, but this is not always the case for the comparator group e.g. some ethnic

[16] This is a categorical variable derived from the highest level of attainment (including whether this was vocational, academic or an apprenticeships), HE subject, degree classification, and HE institution

groups are disproportionately located in particular regions in England. This may be contributing to some of the uncertainty seen in the results.

Some of the percentages for some of the ethnicity comparisons are very high and these appear to happen in two different scenarios which add to the complexity of interpretation:

1. Where the other ethnic group becomes even less likely to be in a good outcome than white British after adding the controls, compared to the observed difference (e.g. Black African compared to white British males – Figure 16 of the ethnicity chapter). For poor outcome this is when the other ethnic group becomes even more likely to be in a poor outcome than white British.
2. Where there is a change of sign in the probit regression after adding the controls e.g. Chinese and white British males and good outcome (observed to be 4.7 percentage points **more** likely to be in good outcome than white British, but 4.5 percentage points **less** likely after adding controls- Figure 16 of the ethnicity chapter).

We are less confident in the results from these particular comparisons although the trends of which factor or groups of factors are important are consistent with what we would expect and adds further weight to the argument that the reasons for differences between labour market outcomes of different ethnic groups are complex, varied and not well understood.

Probit regression with sequential controls (ethnicity only)

Due to uncertainty of the results from the ethnicity decomposition, additional probit regression analysis was carried out which sequentially added groups of control variables to attempt to ascertain how these variables affected good or poor outcome for different ethnic groups and how much of the gap between white British and the other ethnic group each control or group of controls explained.

These regressions use the same methodology as the probit regression outlined previously (Regression analysis) but with some variation in the controls which are used in each model. Five models were run for each of males and females, good and poor outcome.

Equation 6: Ethnicity sequential probit model 1

$$\mathrm{P(Outcome\_Var} = 1) = \Phi(\beta_0 + \beta_1 \mathrm{Ethnicity})$$

Model 1 (Equation 6): no controls added. Here we assume that all variation in good (or poor) outcome is due to ethnicity.

Equation 7: Ethnicity sequential probit model 2

$$\mathrm{P(Outcome\_Var = 1) = \Phi(\beta_0 + \beta_1 Ethnicity + \beta_2 SES)}$$

Model 2 (Equation 7): SES only. Socioeconomic status quintile is the only control.

Equation 8: Ethnicity sequential probit model 3

$$\mathrm{P(Outcome\_Var = 1) = \Phi(\beta_0 + \beta_1 Ethnicity + \beta_2 SES + \beta_3 Demographics)}$$

Model 3 (Equation 8): As model 2 with demographic controls (local authority during KS4 and current residence, EAL, SEN status and season of birth)

Equation 9: Ethnicity sequential probit model 4

$$\mathrm{P(Outcome\_Var = 1) = \Phi(\beta_0 + \beta_1 Ethnicity + \beta_2 SES + \beta_3 Demographics + \beta_4 School)}$$

Model 4 (Equation 9): As model 3 with school factors (KS2 and KS4 attainment, school type, school progress measure, school demographics)

Equation 10: Ethnicity sequential probit model 5

$$\mathrm{P(Outcome\_Var = 1) = \Phi(\beta_0 + \beta_1 Ethnicity + \beta_2 SES + \beta_3 Demographics +}$$

$$\mathrm{\beta_4 School + \beta_5 Post16)}$$

Model 5 (Equation 10): All controls: as model 4 with post-16 education (A levels, further education, higher education and highest level of education). This is our best estimate (using the socioeconomic, demographic and education variables in the administrative data) of the gap in good (or poor) outcome between ethnic groups.

Full results for good and poor outcome, males and females, for the six ethnic groups examined in the ethnicity chapter are available in the accompanying tables (Tables 1.12 to 1.15) but these are not shown in the ethnicity chapter. See Regression analysis for an explanation of how to read these tables.

The regression results begin to show that across the ethnic groups different factors are associated with labour market outcomes. However, it is difficult to assess the importance of each group of factors. This is fairly clear-cut for the comparison of some ethnic groups to white British – such as Indian females (both good and poor outcome), where each additional group of factors explains a proportion of the difference between Indian and white British females. Some are more complex, such as good outcome for black Caribbean females compared to white British females – some factor groups are working to increase the gap between the two ethnic groups and some to decrease the gap.

In addition, these groups of factors are not standalone. For example, those from disadvantaged backgrounds are more likely to have lower attainment at KS2 and KS4 and are less likely to go to university, or attend lower quality higher education institutions. If we were to add the controls in a different order, such as post-16 education first, the proportion of the gap they explain would appear different (because SES, demographics and school factors explain some of the differences in post-16 education). This makes it difficult to determine the true importance of each one.

In combination with the decomposition analysis, the sequential controls regression shows the complexity of the factors involved in labour market outcomes for ethnicity.

Department
for Education

For any enquiries regarding this publication, contact www.education.gov.uk/contactus.

This document is available for download at www.gov.uk/government/publications.