# ABCS: Associations between characteristics of students

## November 2026

## Methodology

á

**Enquiries to** official.statistics@officeforstudents.org.uk

**Publication date** 5 November 2026

# Contents

**Introduction** **2**

**Statistical modelling** **3**

Modelling approach 3
Dealing with combinations not in the data 4
Dealing with unknown values and missing characteristics 5

**Characteristics: data sources and definitions** **7**

Data sources 7

**Using the ABCS measures** **11**

How to obtain ABCS quintiles for individuals or groups of individuals 11
How to interpret the dashboards 11
Further information 11

# Introduction

1. Association between characteristics of students (ABCS) is a set of measures that seeks a better understanding of how outcomes vary for groups of students holding different sets of characteristics. We define groups of students by looking at a set of characteristics so that we can determine the relationship of not just one characteristic on an outcome, but the relationship of multiple characteristics.

2. Each measure takes the form of a set of quintiles, where quintile 1 indicates groups of students who are least likely to experience or report a positive outcome and quintile 5 indicates group that are most likely to experience or report a positive outcome.

3. This document describes the core methodology used in creating all the ABCS measures, and unless otherwise indicated, the information presented here is relevant to all measures. Where possible, we have used a consistent methodology across all ABCS measures.

4. To see the list of ABCS measures and details about each of them, please refer to the documents on the ABCS page of our website.[1]

[1] See Associations between characteristics of students - Office for Students.

# Statistical modelling

## Modelling approach

5. For each ABCS measure, we use a statistical model to predict how likely an individual is to have a positive outcome based on the characteristics included in the model. The use of statistical modelling tells us whether there is a statistically significant relationship between each characteristic used and the outcome, while controlling for the impact of the other characteristics in the model. In effect, it's like saying "if every individual in this model had exactly the same set of characteristics except for this one characteristic I'm looking at, would there be a relationship between this characteristic and having a positive outcome?". Modelling is also useful where there are only a very small number of individuals with a particular characteristic. When this happens, it's difficult to say whether the observed behaviour is a trend that is likely to be true of anyone with that characteristic, or if the observed behaviour is atypical and just happens to have been seen in the individuals you have. The use of statistical modelling gives us a 'best estimate' of the likely outcome of people holding those characteristics, based not only on the observed outcomes, but also accounting for the behaviour of those holding some of the same characteristics.

6. Where possible, each model uses the five most recent cohorts of data available at the time of development to maximise the amount of information the statistical models have to generate their predictions. Where five years of data are not available, we use as many years as we can.

7. The ABCS modelling methodology uses a binary logistic regression model with all characteristics included as main effects. We then test every two-way interaction to assess whether it improves the fit of the model.[2] Those that do are included in the final model.

8. Higher order interactions were also considered when developing the modelling approach for our ABCS measures. However, these higher level interactions led to the models becoming unstable, meaning that the predictions were not reliable.

9. Having derived the predicted probabilities of a positive outcome for an ABCS measure, we used these to rank the students included in the modelling. We then took the rate at the boundary of each 20$^{th}$ percentile; these five values became our 'breakpoints', which define the lower boundaries of our quintiles for this measure. Subsequently, given a student group defined by a combination of characteristics, we assign it a quintile by comparing the predicted probability of that group with the breakpoints; at the boundary between two breakpoints, if the rate for the group is less than the breakpoint it will fall into the lower quintile; if it falls exactly on or above the breakpoint it will fall into the higher quintile. Using this approach, two groups with the same predicted probability will always be assigned the same quintile; however, this means that each quintile may not contain exactly 20 per cent of the population.

10. In selecting the characteristics for use in each of the models we were looking for characteristics that should not influence the outcome in question, but where there was evidence that the outcomes for groups within these characteristics differed. For example, there is no reason why

[2] Note that for our older models, this was done using a stepwise approach with an entry and stay criterion of $\alpha$=0.05. For models developed in 2026 or later, this was determined by comparing the AICs of models with various combination of interactions terms included.

a student's ethnicity should have an impact on the likelihood of them continuing into the second year of their course. However, our analysis of continuation rates showed that black students had lower continuation rates than students from any other ethnic background.[3] Conversely, while we know that prior attainment will have an impact on the likelihood of a young person entering higher education, this has not been included in the model because this is a justifiable – or valid – relationship.

11. The data quality framework was used to guide our selection of characteristics.[4] No characteristics were included if data was not available for at least half of the cohorts being modelled. Although it would have been possible to include more characteristics than we have done (evidenced by their having an observable relationship with the outcome), having too many characteristics would potentially destabilise the model results due to overfitting.

12. During characteristic selection we also considered multicollinearity: this occurs when two or more characteristics are highly related to each other as well as the outcome and it can cause model instability.

13. In instances where data for a characteristic is missing, unknown or not applicable, we include them as their own 'Unknown or N/A' group in the modelling. Examples where data is not applicable include a TUNDRA (tracking underrepresentation by area) quintile for a student 21 years of age or older, an Adult HE quintile for a student under 21 years of age (see Table 2 for details about these characteristics), or an IMD (Index of Multiple Deprivation) or TUNDRA quintile for one domiciled outside of England.

14. For some characteristics, there are groups that contain very small numbers of students. Where possible, we have kept them as their own distinct group for use in the modelling, but where their numbers are too small to maintain model stability, we have combined them with the group having the closest outcome rate.

## Dealing with combinations not in the data

15. Although the modelling data used for an ABCS measure consists of several cohorts of actual student data, it will not necessarily contain every unique combination of characteristic values possible. For example, there may be no Asian or Asian British - Bangladeshi students who are also distance learners in the modelled data. As the intention is that ABCS quintiles can be applied to future student data, or other sources of data, this would cause an issue if no quintile was assigned to such groups. Therefore, we derive ABCS quintiles for these groups by calculating the predicted probabilities using the model's parameter estimates, and assign quintiles based on the calculated breakpoints. These new combinations do not contribute to calculation of the breakpoints. Using this method allows us to cater for any combination of characteristics as required.

[3] See Access and participation data dashboard - Office for Students.

[4] See Differences in student outcomes - further characteristics - Office for Students (Annex A: Data quality framework).

## Dealing with unknown values and missing characteristics

16. For most of the characteristics used in the modelling, there are student groups with unknown values. In the modelling data we have not differentiated between values that are unknown for structural reasons (e.g. the data may not be collected by that provider type) and values that are unknown for other reasons (e.g. an invalid postcode resulting in unknown geographical measures). For each characteristic they are grouped into the category 'Unknown or N/A'.

17. When applying ABCS quintiles, we have decided not to assign these characteristics the value 'Unknown or N/A', as we do in the modelling data. Instead, we calculate an average of the predicted outcome rates associated with the known values for that characteristic, weighted by the number of students in each of these groups.

18. Table 1 gives an example of what would happen if our ABCS measure had only one characteristic. From the modelled data we can see the number of students in each of the three free school meal (FSM) eligibility groups, the predicted probabilities calculated from the model outputs and the quintile they are assigned to. To assign a quintile to a student whose FSM eligibility status is unknown, we calculate a weighted average of the outcome rates across the three groups. Then, using the breakpoints, we derive the corresponding ABCS quintile.

19. The calculation of the weighted average for the missing data is as follows:

    ((700 x 0.85) + (250 x 0.45) + (50 x 0.59)) ÷ 1000 = 0.737

    That is, the sum of the number of students in each group multiplied by their predicted outcome rate divided by the total number of students.

20. This rate can then be compared with the breakpoints to derive which quintile the group falls into.

**Table 1: Example treatment of unknown FSM eligibility for a student group**

| FSM eligibility status | Number of students | Predicted outcome rate | ABCS quintile | Proportion of students (per cent) |
|---|---|---|---|---|
| Not eligible for FSM | 700 | 0.85 | Quintile 5 | 70.0% |
| Eligible for FSM | 250 | 0.45 | Quintile 2 | 25.0% |
| Unknown or N/A | 50 | 0.59 | Quintile 3 | 5.0% |
| Missing data | | 0.74 | Quintile 4 | |

21. Although the example in Table 1 is a significant simplification, the same process can be applied when there are multiple characteristics, using the model outputs to calculate the predicted outcome rate for any combination of characteristics and calculating the weighted average for all groups of interest. This same weighted average approach is extended so that ABCS quintiles can be derived for cases where any number of characteristics have missing or unknown values. It must be noted however, that the more characteristics that have missing or unknown

values, the less confidence can be placed on the derived quintile so they should be used with caution in these cases.

# Characteristics: data sources and definitions

22. This section summarises all the characteristics that are used across the ABCS measures. Each measure uses a different subset of these characteristics. For some measures, groups within a characteristic will have been combined due to small issues causing problems with the statistical model. For details of where this has been done, see the documentation for each individual ABCS measure.[5]

## Data sources

23. The following data sources are used in the derivation of ABCS measures:

    a. The Department for Education's (DfE) National Pupil Database (NPD) this contains data on school pupils in England.[6]

    b. The designated data body's (DDB) Student Record, and student and student alternative records where data is needed for pre 2022-23.

    c. The DfE's Individual Learner Record (ILR).

    d. Data from the Student Loans Company (SLC).

    e. The Graduate Outcomes Survey.[7]

    f. The National Student Survey (NSS).[8]

24. Additional geographically based measures are also used, details of the data sources for these are covered in the table below. All the characteristics used in the ABCS measures are in listed in Table 2, alongside their data source and a brief description. Further information on these characteristics can be found in our student outcome and experience measures core algorithms document.[9] The field name used in the core algorithms document is in brackets after the characteristic.

[5] Available from Associations between characteristics of students - Office for Students.

[6] The DfE does not accept responsibility for any inferences or conclusions derived from the National Pupil Database (NPD) data by third parties.

[7] See Thousands of graduate voices. One survey - Graduate Outcomes.

[8] See National Student Survey: National Student Survey.

[9] See Documents describing our measures and definitions - Office for Students (Technical algorithms for student outcome and experience measures).

**Table 2: Characteristics used in ABCS measures**

| Characteristic (field name) | Data source | Description |
|---|---|---|
| **Adult higher education quintile (adult_he_quintile)** | OfS | Assigned to an area based on the proportion of adults that held a higher education at the 2011 census. ABCS models restrict this to those aged 21 or over at the start of their course. |
| **Age (engagement_starting_age)** | DDB Student record, ILR | The age of the student on the 31 August in the year in which they started their course. Ages are grouped differently across the ABCS measures – see measure specific reports for details. |
| **Care experience (care_leaver_status)** | DDB Student record | Whether a student has ever spent time in the care of a local authority in England or Wales, or Health and Social Care Trust in Northern Ireland, or who has self-declared as being in care for three months or more.[10] |
| **Disability (reported_disability_type)** | DDB Student record, ILR | Disability as self-reported by students. |
| **Estrangement (ever_been_estranged)** | SLC | Whether a student has ever reported being estranged from their parents. |
| **Ethnicity (student_ethnicity)** | DDB Student record, ILR, NPD[11] | Ethnicity as self-reported by students. |

[10] Detailed definitions are provided here: EntryProfile_CARELEAVER 25056 | HESA

[11] Note that the categories differ between the NPD record and the OfS core algorithms. See Find and explore data in the National Pupil Database - GOV.UK for details.

| Characteristic (field name) | Data source | Description |
|---|---|---|
| **FSM eligibility**[12] | NPD | Whether the student was ever recorded as being eligible to receive free school meals in the six years prior to the March census date in their final year of key stage four (year 11). |
| **Gender (NPD only**[13]**)** | NPD | Records gender of school pupils. |
| **Income Deprivation Affecting Children Index (IDACI) (idaci_quintile)** | Ministry of Housing, Communities and Local Government (MHCLG) | Measures the proportion of children under the age of 16 in low-income households for a particular area in England.[14] The most deprived areas are in quintile 1 and the least deprived are in quintile 5. |
| **Index of Multiple Deprivation (IMD) (home_imd_quintile_by_nation**[15]**)** | Ministry of Housing, Communities and Local Government (MHCLG) | Measures levels of deprivation for a small area within England.[16] We use IMD quintiles, where the most deprived areas are in quintile 1 and the least deprived are in quintile 5. |
| **Local or distance learner (study_location_type)** | DDB Student record, ILR | Whether a student's study location is considered to be local to their home address or if they are a distance learner. |
| **National Statistics Socio-Economic Classification (NS-SEC) (socioeconomic_class)** | DDB Student Record | Records the classification of employment of the highest earner in a student's household. For students aged under 21 at the start of their course, it is based on the occupation of their highest earning parent. For older students, it is based on their occupation prior to starting their course. |

[12] see "Fields used for determining students' eligibility for free school meals at key stage 4" section of core algorithms

[13] See Find and explore data in the National Pupil Database - GOV.UK

[14] See English indices of deprivation 2019 - GOV.UK.

[15] Note that this field also includes the equivalent measures for Scotland, Wales and Northern Ireland, but only English IMD is used in the ABCS measures.

[16] See English indices of deprivation 2019 - GOV.UK.

| Characteristic (field name) | Data source | Description |
|---|---|---|
| **Parental higher education (parental_education)** | DDB Student Record | Records the response to the question 'Do any of your parents have any higher education qualifications, such as a degree, diploma or certificate of higher education?' where a parent is defined as 'natural parents, adoptive parents, step-parents or guardians who have brought you up.' |
| **Religious belief (religious_belief)** | DDB Student Record | Religious belief as self-reported by the student. |
| **Sex (student_sex)** | DDB Student Record, ILR | Sex as self-reported by the student. |
| **Sexual orientation (sexual_orientation)** | DDB Student Record | Sexual orientation as self-reported by the student. |
| **TUNDRA (tracking underrepresentation by area) (tundra_msoa_quintile)** | OfS | An area-based measure that uses tracking of state-funded mainstream school pupils in England to calculate young participation.[17] Quintile 1 areas have the lowest levels of participation and quintile 5 areas have the highest participation. It only applies to students aged under 21 at the start of their course. |

[17] See About the TUNDRA area-based measures data - Office for Students.

# Using the ABCS measures

## How to obtain ABCS quintiles for individuals or groups of individuals

25. To obtain the quintile for a small number of individuals or groups of individuals with exactly the same set of characteristics, the ABCS dashboards can be used.[18] The categories for each characteristic can be chosen from the dropdown menus, and a derived quintile will be shown. If the individual or groups in the data you are using has a characteristic value that is missing, then the dropdown can be left as '—any [characteristic]—' and the derived quintile will reflect the weighted average approach as outlined in paragraphs 16 to 20.

26. To obtain quintiles for multiple individuals, use the data downloads to look up their quintiles based on the individuals' characteristics.[19] If any attribute groups have been combined in the modelling data (for example, small groups such as the Gypsy or Traveller ethnic group are often combined with a larger ethnic group for ABCS modelling), then any individuals with this attribute should have their category changed to match the larger group. This will allow the individual to be matched with the ABCS data. Information about which categories have been combined in this way can be found in the report associated with each ABCS measure.

## How to interpret the dashboards

27. There is a separate dashboard for each ABCS measure. Using the dropdown menus, categories can be selected for each characteristic, or '—any [characteristic]—' can be selected instead of choosing a particular attribute.

28. The lower panel shows the distribution of individuals across the ABCS quintiles. If all characteristics have an attribute selected, then 100 per cent of individuals will fall into a single quintile. This is because a group is defined by a combination of attributes, and each group has a single quintile.

29. If one or more characteristics are set to '—any [characteristic]—then the lower panel will show what proportions of students with any combination of those characteristics fall into each quintile. In addition, at the bottom of the dashboard a derived quintile is shown. This reflects a weighted average of the distribution of individuals across all quintiles, as described previously in this report.

## Further information

30. This series of reports provides an overview of the methodology developed for ABCS measures. For accessibility, high level descriptions have been used, rather than details of the implementation of the methodology in our software environment. If you have any questions, please get in touch: contact details are provided on the front page of this report.

[18] See Associations between characteristics of students - Office for Students

© The Office for Students copyright 2026

This publication is available under the Open Government Licence 3.0 except where it indicates that the copyright for images or text is owned elsewhere.

www.nationalarchives.gov.uk/doc/open-government-licence/version/3/