Associations between characteristics of students

How do outcomes differ when accounting for multiple student characteristics?

Reference OfS 2019.34

Enquiries to Annalise Ruck at official.statistics@officeforstudents.org.uk

Publication date 26 September 2019

Contents

Summary

1. ‘Association between characteristics of students’ (ABCS) is a set of analyses that seeks to better understand how outcomes vary for groups of students holding different sets of characteristics. Higher education outcomes include access to higher education, continuation in higher education, degree attainment and employment outcomes. We define groups of students by looking at a set of characteristics so that we can determine the effect of not just one characteristic on an outcome, but the effect of multiple characteristics. This report focuses on the outcomes of entry to higher education and continuation in higher education, but the ABCS work sets out a framework that could be used for analysis of other student outcomes, such as degree attainment and employment after graduation.

2. This report sets out the methodology for creating ABCS measures and gives specific details of how we have created the ABCS continuation and ABCS access measures. We discuss examples of where these measures have shown groups of students with poor outcomes that may not have been seen when only looking at single characteristics.

3. This is an experimental official statistic. As such, we are looking for feedback regarding any improvements that could be made to the methodology or the presentation of the ABCS measure. We are also keen to understand how you might use these measures. Please get in touch with us to let us know your thoughts and feedback.

Why ABCS?

4. At the Office for Students we want every student, whatever their background, to have a fulfilling experience of higher education that enriches their lives and careers. One of the ways in which we can assess how well we're meeting that aim is by carrying out analysis that looks at student outcomes by different characteristics.

5. There are many outcomes across the student lifecycle, but most commonly we look at access to higher education, continuation in higher education, degree classification, and employment or further study. The characteristics we look at tell us something about the student or their background, such as their ethnicity or sex. Typically, our analyses have looked at student characteristics one at a time, comparing categories within the characteristics to see how the outcome differs between them. For example, looking at access rates by sex and comparing men and women.

6. However, single-characteristic analysis is an over-simplified view of the impact that characteristics have on an outcome. None of us has only a single characteristic that makes up who we are, but a great many different characteristics. It is possible that the combination of these characteristics can result in different outcomes than when looking at a single characteristic. Where we have begun to look at more than one characteristic, we can already demonstrate that we find very different outcomes to what is observed when looking at a characteristic in isolation.

7. For example, our analysis on continuation rates shows that, when looking at all students together, continuation rates are highest for students from POLAR4[1] quintile 5 areas, and lowest for those from quintile 1 areas. However, if we look at continuation rates for black students by POLAR4 quintiles, we find that continuation rates are highest for those from quintile 1 areas. This is important because, with an understanding that quintile 1 students are the least likely to continue, efforts to improve continuation rates will be focused on quintile 1 students. However, this will mean that amongst black students, the groups with the lowest continuation will miss out on any intervention.

8. UCAS has already developed a measure which looks at multiple characteristics and how they impact access to higher education: the multiple equality measure (MEM)[2]. The UCAS MEM calculates the probability of entry to higher education aged 18 based on a range of equality characteristics and their combinations. This analysis has drawn heavily on the methodology used for the MEM to allow us to create a framework for developing similar measures for outcomes across the student lifecycle. In this report, we focus on using this framework for creating a similar measure for continuation rates, but a comparison with the MEM methodology is also provided to allow us to understand the consistency of the two approaches.

9. Of course, it is not possible to capture everything about an individual in order to predict their likely outcomes, and this is not our aim. We want to identify groups of students who are most

[1] For details about POLAR4 visit: www.officeforstudents.org.uk/data-and-analysis/young-participation-by-area/

[2] See https://www.ucas.com/data-and-analysis/ucas-undergraduate-releases/ucas-undergraduate-analysis-reports/equality-and-entry-rates-data-explorers

likely to be at risk of having poor outcomes by considering several different characteristics. This allows us to achieve a clearer picture of which groups of students should be targeted and monitored in order to close gaps in outcomes between different groups of students.

10. The exploration and development of experimental measures for access and continuation that depend on the relationship between multiple student characteristics builds on the approach set out in the OfS's reforms to access and participation, on which we consulted last year. Our current guidance encourages providers to consider and address combinations of characteristics through their self-assessment and where appropriate their targets, outreach and student support. It is our hope that, in the future, measures such as the experimental ones described here will help providers to better address combinations of characteristics when seeking to understand how access, student success and progression into further study and work vary across different groups of students. This, in turn, should lead to more effective targeting and evaluation of interventions and help continue to close equality gaps across the student lifecycle.

11. In this report we set out a methodology for developing measures of outcome that categorise groups of students by the likelihood of them achieving the outcome based on a set of characteristics. These outcome groups allow easy identification of groups of students who are most at risk of having poor outcomes. In the case of access, those in access group 1 are the least likely to enter higher education, whilst those in access group 5 are the most likely to enter higher education. Likewise, those in continuation group 1 are the least likely to continue into the second year of a full-time undergraduate course, whilst continuation group 4 are the most likely to continue.

12. Creating groups in this way makes it much easier to identify groups of students who are most at risk of having poor outcomes. The user need only look at the groups of student within the lowest outcome groups, rather than looking at all possible groups of students with every possible combination of characteristics.

13. It is our intention that this more complete understanding of the kinds of students who are, at present, less likely to achieve these outcomes can be used by higher education providers to effectively target groups of student, and by the OfS to monitor gaps in outcomes for these same student groups.

14. In this report, we focus specifically on the outcomes of continuation in and access to higher education, but this framework has been created in such a way that it can be adapted and used for the creation of measures for other outcomes such as degree attainment and employment outcomes.

15. The report sets out the methodology for creating ABCS measures – a framework that can be applied to many different outcomes. We have first used it to develop an ABCS continuation measure, and so details of this are set out first. Secondly, we have used the methodology to create an ABCS access measure. This has allowed us to compare our methodology with UCAS' MEM methodology. The continuation and access sections of the report are quite repetitive, since they cover a similar methodology, but are designed to stand alone.

These measures are being published as experimental. We are seeking feedback on all aspects of the methodology and the measures.

If you have any comments, criticisms or suggestions, or if you require any further information or details, please contact Annalise Ruck at official.statistics@officeforstudents.org.uk.

Methodology

16. We have created this methodology to be applicable to a range of outcomes. For each outcome, the method is applied independently, creating a separate set of outcome groups in each case. In this report, we create outcome groups for access to higher education and continuation into the second year of full-time undergraduate. It is likely that in the future we will also create groups for continuation in part-time courses, degree attainment and employment after graduation.

17. For each group of students, defined by every possible combination of the characteristics we are using, we calculate modelled outcome rate. This is likelihood of that group of students achieving an outcome based on statistical modelling. For the statistical model, we use data for five previous cohorts of undergraduate degree[3] students, in the case of continuation, and school pupils in the case of access. For these models, we use individualised data from the Department for Education's (DFE's) National Pupil Database[4], the Education and Skills Funding Agency's (ESFA's) Individualised Learner Record (ILR) and the Higher Education Statistics Authority's (HESA's) student record and alternative provider student record. In each case, we use five cohorts of data in order to ensure that there are enough students in each group to be able to carry out statistical modelling.

18. For each outcome the modelled rates are used to create a set of outcome groups that indicate how likely it is for a group of students to achieve that outcome, where students in the lowest outcome groups are the least likely to achieve that outcome and those in the highest group are the most likely to achieve that outcome. We define group membership by the modelled rate of achieving that outcome for a group of students with a specific set of characteristics. For example, if a group of students has a very low modelled rate of entering higher education, they will be in access group 1.

19. We have also taken a statistical approach to define the outcome groups. Groups are created in such a way that there is as clear a differentiation between each group as possible, while also trying to keep the number of students in each group as similar as possible. This means that there is assurance that the modelled rate for a group of students in one outcome group really is different from that of a group of students in the outcome group above or below. Details of how we have done this are in the 'Creating outcome groups' section below.

Modelling approach

20. In order to calculate the modelled rates for groups of students, we employ a statistical modelling approach. Use of statistical modelling allows for assessment of whether there is a statistically significant relationship between the characteristics used and the outcome. Additionally, in the case of smaller student groups, it is not always safe to assume that the recorded behaviour of people in that group would reflect the behaviour of a larger group of people holding those same characteristics. The use of statistical modelling gives us a 'best

[3] Often referred to as first degree students.

[4] Contains data sourced from the DfE's National Pupil Database. The DfE does not accept responsibility for any inferences or conclusions derived from the NPD data by third parties.

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.

21. Since the two outcomes here are binary, that is, they have two possible values: achieving the outcome or not achieving it, we use a binary logistic regression model. The models calculate the modelled rate of achieving the outcome using the characteristics we have chosen to use (known as the factors). A stepwise selection method is used with an entry and stay criteria of $\alpha=0.05$. All main effects are kept in the model.

Choosing the factors

22. In selecting the factors for use in these models we are looking for characteristics that should not influence the outcome in question, but where there is evidence that the outcomes for groups within these characteristics differ. For example, there is no reason why 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 shows that black students have lower continuation rates than students from any other ethnic background. Conversely, whilst we know that prior attainment will have an impact on the likelihood of a young person entering higher education, this will not be included in the model because this is a justifiable – or valid – relationship.

23. As well as looking at characteristics that we have included in previous analysis, for this model we have also considered other personal characteristics and area-based measures. In each case, we have explored the relationship between that characteristic and the outcome in question before including it in the model. The final models only contain those factors that have been found to be statistically significant.

24. The statistical models for the different outcomes will not necessarily contain the same factors. It is likely that there will be overlap between the different models, since there are many characteristics which are related to all outcomes, and there are some characteristics which are related to some outcomes, but not others. For this reason, we have undertaken exploratory analysis to determine factor selection separately for each outcome.

Choosing the interactions

25. In order to allow the model to calculate the best estimates of the outcome rates we test both main effects and interaction effects. Interactions are not necessarily included for all possible values within a characteristic. For example, if we consider the characteristics of age and sex, it is possible that the interaction of being 18 and female is included, but the interaction between being 51+ and female might not be included. To determine which interactions to include, we create dummy variables for every possible combination of values from two characteristics. These dummy variables are then included in the model alongside the main effects and the stepwise selection method is used to determine which interactions to keep in the model.

26. Only two-way interactions are included in the model. We considered higher order interactions as part of the preliminary analysis, but the number of possible factors created led to the model becoming unstable. This means that the estimates that are calculated become unreliable.

Creating outcome groups

27. We have designed the methodology for creating the outcome groups with the following principles in mind:

    a. Groups should only be split where there is a clear differentiation between the modelled outcome rates of the groups.

    b. Groups should be kept as similar in size as possible. Groups should not represent less than five per cent of the student population.

    c. The preference is for there to be five risk groups.

    These principles are hierarchical in that part a. is the most important and part c. the least.

28. Our desire to have five groups is based, in part, on the fact that we report other measures in quintiles, such as the Index of Multiple Deprivation (IMD) or the Participation of Local Areas (POLAR4) measure. Therefore, having five groups would be useful for the sake of comparison. However, since we are most concerned that there is clear differentiation between the groups, we recognise that there might be little variation in the likelihood of an outcome in the central groups, with only distinct groups identifiable at the highest and lowest levels. If this turns out to be the case, it will be preferable to have fewer clear groups, rather than five. Likewise, if we find that creating clearly differentiable groups leads to very unequal outcome groups in terms of the number of students captured in groups, we will consider using that grouping provided we are able to clearly identify those in the lowest outcome group.

29. Groups of students with a combination of characteristics that are held by fewer than 50 students are not used in the creation of the outcome groups, although they are put into the outcome groups once they have been defined. This is because small groups are likely to have high levels of uncertainty around their modelled outcome rate. This means that where two groups' modelled rates are very different, they might, in fact, be very similar when accounting for uncertainty. Removing these small groups of students prevents the choice of boundaries for the outcome groups being unduly influenced by these small groups.

30. We order the remaining groups by the size of their modelled rates of achieving the outcome before calculating the difference between each modelled rate. We select the largest differences as potential 'breakpoints' – that is, a point at which to split the data to create a group. Depending on the number of outcome groups we are choosing to create, we use statistical methods to select the required number of breakpoints which maximise the equality of the group sizes. Since these are the differences between two modelled rates, we use the student group with the largest modelled rate of the two to define the lower boundary of the outcome group.

31. Once we have defined the boundaries, we put the small student groups into the outcome groups based on their modelled rates. Where the modelled access rate for a small group of students is between the group boundaries, we put them into the lower group (apart from rates that are below the lower boundary for group 1 – which we include in outcome group 1; or rates that exceed the upper boundary of group 5 – which we include in outcome group 5).

32. Where the modelled rates are very close it can be difficult to find breakpoints which give outcome groups that meet the first principle. In this case, we may choose to create fewer

groups to ensure that there are clear differences in the modelled rates between the outcome groups.

33. We use this method in preference to quintiles because of the focus on the differentiation between the groups, rather than the size of the groups. This is because of our proposed use of the measure to target and monitor groups of students who are clearly less advantaged than others in relation to a given outcome.

Looking at continuation

34. The first outcome to which we have applied the ABCS framework is continuation for full-time students. Full-time continuation is measured one year and 14 days after a student starts their studies. Students are defined as continuing if they are continuing with or have completed their studies, or have transferred to another higher education provider to continue their studies[5].

35. We have used continuation data for UK-domiciled undergraduate students who started their courses between the academic years 2012-13 and 2016-17 and were studying full-time at an English provider. We have used data from the ESFA individualised learner record (ILR) and the HESA student record and alternative provider student record. Combining data from five cohorts allows us to carry out robust analysis, ensuring that there are enough students in each of the characteristic groups to allow us to carry out analysis regarding their continuation behaviour.

36. We create four continuation groups, where groups of students in continuation group 1 are the least likely to be continuing their studies and those in continuation group 4 are the most likely to be continuing.

Choosing factors to include in the model

37. Previous analysis[6] we have undertaken has found differences in continuation rates for student groups within the following characteristics: age, disability, POLAR4, IMD, ethnicity and sex. Therefore, we have included all of these characteristics in the initial modelling. Further investigation of the relationship between age and continuation has shown that there are differences in continuation rates between those who start their course aged 18, 19 or 20, whilst there is little difference in rates between those aged 25 and 50. This has led us to include a different grouping of age than has been used in previous publications. Similarly, we have investigated whether differences in continuation rates exist at a lower level of ethnicity than the five groups that are usually used. Looking at ethnicity split into 18 levels has shown that continuation rates differ substantially between ethnicities within the broad categories of Asian, black and white. Therefore, we have chosen to include ethnicity at this more detailed level; details of this grouping can be found in Annex A[7].

38. Anecdotal evidence suggests that there are differences in continuation rates between local or distance learners and those who are neither local nor distance learners. Because of this, we have investigated the relationship between the locality of a student and continuation. Students are grouped as either local or distance learners or not local or distance learners[8]. Exploratory analysis shows that the average continuation rate for local or distance learners across the

[5] For details of how we calculate continuation measures from the data, see the OfS access and participation data methodology and rebuild instructions: www.officeforstudents.org.uk/data-and-analysis/access-and-participation-data-dashboard/guide-to-the-access-and-participation-data-resources/

[6] See www.officeforstudents.org.uk/data-and-analysis/continuation-and-transfer-rates/

[7] Available at www.officeforstudents.org.uk/publications/associations-between-characteristics-of-students/

[8] For details of the variables used to determine whether a student is a local or distance learner, B3MONLOCAL, see the OfS core algorithms document, available at: www.officeforstudents.org.uk/data-and-analysis/access-and-participation-data-dashboard/guide-to-the-access-and-participation-data-resources/

academic years 2012-13 to 2016-17 is 86.5 per cent, compared to 92.2 per cent for non-local or distance learners. Since this is a large difference, we have included the local or distance learner marker in the initial model.

39. Alongside POLAR4 and IMD, we have considered a third area-based measure: the income deprivation affecting children index (IDACI)[9]. This measures the proportion of children under the age of 16 in low income households for an area. It is calculated at lower-layer super output area (LSOA) level and is a supplementary measure to IMD. The inclusion of IDACI allows us to further understand the area that the young person comes from. Inclusion of IDACI means that there are three area-based measures in the model: IDACI, IMD and POLAR4. We will examine the relationship between these variables, and whether this leads to multicollinearity in the model, as part of the modelling process.

The model

40. We have used a binary logistic regression model to calculate the modelled rate of a group of students continuing in their studies. A stepwise selection method has been used with an entry criterion of α=.05 and a stay criterion of α=.05. The factors included in the final model are: age, disability, ethnicity, IDACI, IMD[10], local or distance learner, POLAR4[11] and sex. Having established that all these factors remain in the model using the stepwise selection method, we consider two-level interaction. Every possible combination of values from two characteristics are considered for entry into the model as two-way interaction effects. Those which meet the stepwise selection criteria remain in the model. Details of the categories within each of the variables and the two-way interactions that are included can be found in Annex A and Annex B respectively[12].

41. Therefore, the final model is:

$$logit(\pi_i) = \beta_0 + \beta_1 age_i + \beta_2 disability_i + \beta_3 ethnicity_i + \beta_4 IDACI_i + \beta_5 IMD_i + \beta_6 local\ learner_i + \beta_7 POLAR4_i + \beta_8 sex_i + interactions + \epsilon$$

where i is an individual.

The grouping

42. The model generates modelled rates of continuation for 43,151 different groups of students, based on the combination of the eight characteristics and measures included in the model. The distribution of these modelled rates is not particularly spread, with some notably low and high estimates, but the majority being very similar. We have calculated five continuation groups, but concluded that the middle group is too small, representing only five per cent of the students

[9] See https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015

[10] Because IMD and IDACI are only calculated for areas in England, student who are domiciled outside of England are assigned a value of 'N/A' for these two factors.

[11] Because POLAR4 is only used for students who were under 21 at the start of their course, students who were 21 or over at the start of their course are assigned a value of 'mature' for POLAR4.

[12] Available at: www.officeforstudents.org.uk/publications/associations-between-characteristics-of-students/

included in the model. Therefore, we have chosen to have four continuation groups, where continuation group 1 has the lowest modelled continuation rates and continuation group 4 has the highest modelled continuation rates. The groups are as follows[13]:

a. Continuation group 1 contains groups of students with modelled continuation rates between 58.4 and 81.2 per cent. This represents six per cent of the students included in the modelling.

b. Continuation group 2 contains groups of students with modelled continuation rates between 81.2 and 84.8 per cent. This represents seven per cent of the students included in the modelling.

c. Continuation group 3 contains groups of students with modelled continuation rates between 84.8 and 96.6 per cent. This represents 82 per cent of the students included in the modelling.

d. Continuation group 4 contains groups of students with modelled continuation rates between 96.6 and 98.5 per cent. This represents five per cent of the students included in the modelling.

43. Clearly, the differences in modelled continuation rates between the groups are very small (into the third of fourth decimal place). Given that these have been defined by a method that maximises the difference between groups, this reflects how narrow the range of modelled continuation rates is. While this results in the vast majority of students being in a single continuation group, it also clearly defines the groups of students at the extreme ends of the range, meaning that we can be confident that those in continuation group 1 really are those most at risk of not continuing, and those in continuation group 4 really are the most likely to continue.

Sensitivity analyses

44. Because the analysis uses five cohorts combined, it is important for us to understand whether there are any big changes in continuation rates in any groups between the cohorts, and how this might impact our ABCS continuation measure. For age, IDACI, IMD, local or distance learner, POLAR4 and sex, there has been little change in continuation rates across the five cohorts. There are small changes in continuation rates for disability types, with students reporting a mental health condition seeing increasing continuation rates over time[14] (from 84.9 per cent for those starting their course in 2012-13 to 86.8 per cent for those starting their course in in 2016-17) and those reporting social or communication impairments seeing a slight decrease, from 90.5 per cent for those starting their course in 2012-13 to 88.1 per cent for those starting their course in 2016-17.

[13] The modelled continuation rates are only given for groups with 50 or more students. The continuation group might contain rates that are outside the given ranges after smaller groups have been entered.

[14] In recent years, we have seen a sharp increase in the proportion of students reporting having a mental health condition. This change is likely to be due to a changes in reporting behaviour and is unlikely to reflect the true change in the proportion of student with a mental health condition. As such, we can't know whether these changing rates are as a result of changes in reporting or actual changes.

45. The biggest differences in rates across the five cohorts is in ethnicities, where several ethnic groups see different rates across the years, and these do not seem to follow any pattern. However, this is observed in the smaller ethnic groups, such as Gypsy, Roma or Traveller, white–Irish and those of unknown ethnicity. As such, we conclude there are no discernible trends in the data and are happy to include the five years of ethnicity data.

46. Since there are no single cohorts that have atypical continuation rates we are happy that the five cohorts can be combined without worry that any year will have an undue influence of the results.

47. As well as looking at the relationship between continuation and the factors included in the model, it is also necessary to consider the relationships between the factors (i.e. test for multicollinearity). This is because strong relationships between factors can lead to instability in the model coefficients, causing us to question the modelled continuation rates. In particular, we had concerns about the relationships between the three area-based measures. To test for multicollinearity, we have looked at correlations between the factors. This showed that there is a strong correlation between IDACI and IMD (ρ=0.883). However, when looking at the variance inflation factor (VIF) and tolerance for IDACI and IMD, having run a regression model including all these factors, there is no evidence of multicollinearity between these two factors[15].

48. In developing the statistical model, we have looked at two different selection methods: forward selection and stepwise selection. In each case, we have tried a variety of entry criteria, including α=0.1, α=0.05, α=0.01 and α=0.001. Both selection methods resulted in similar outputs with similar model fit statistics, although the stepwise selection method resulted in fewer two-way interactions being included, and the forward selection method resulted in some interactions being included that were not statistically significant and did not have meaningful estimated coefficients. For this reason, we have chosen to use the stepwise selection method.

49. Having tested various entry and stay criteria, both have been set at α=0.05. For a dataset as large as this, we felt that a value of 0.10 was too liberal and ran a high risk of leading to overfit[16] in the model. Whilst the size of the dataset might usually lead us to conclude that smaller entry and stay criteria would be more appropriate, this is not the case here. This is because we had hoped to include much higher order interaction terms in the model in order to give the model the best chance of robust estimation of the underlying continuation rate, but have not been able to do so due to the very high number of possible higher level interaction terms. We have mitigated this, in part, by selecting more generous entry and stay criteria: α=0.05. Allowing these less significant terms into the model means they are likely to be acting as proxies for some of the higher-level interactions.

Findings from the continuation groups

50. Having created the continuation groups, we not only investigate which groups of students fall into which continuation groups, but also how much information you need to know about a student in order to determine which continuation group they are likely to be in. When looking at

[15] The VIF for IMD is 4.64 and for IDACI is 4.69.

[16] Overfit occurs when a model is too complex and begins to explain random error in the data rather than the relationship between factors.

only one characteristic, we find there is no single characteristic for which 100 per cent of students with that characteristic fall into a single continuation group. Therefore, in order to understand which groups of students are the most at risk of not continuing their course, or who have very high continuation rates, we need to look at two or more characteristics.

51. When looking at two characteristics, we find:

    a. Of all students who were 51 or over at the start of their course, 29 per cent are in continuation group 1. However, when looking at students who were 51 or over at the start of their course and reported having a mental health condition, 91 per cent are in continuation group 1.

    b. 42 per cent of all students whose ethnicity is Asian or Asian British–Chinese are in continuation group 4. When filtered to only those who are 18 or under at the start of their course, 72 per cent are in continuation group 4.

    c. Only 11 per cent of female students aged 21-25 are in continuation group 1, a much lower proportion than the 35 per cent of male students aged 21-25 who are in continuation group 1.

52. When looking at three characteristics, we find:

    a. 79 per cent of students of black or black British–Caribbean ethnicity who are aged 21-25 are in continuation group 1, but this increases to 100 per cent when looking at students who have also reported having a mental health condition.

    b. Of all students from a black or black British–African background who are aged 21-25, 50 per cent are in continuation group 1. When this is restricted to those who are local or distance learners, this increase to 91 per cent.

    c. When looking at female Asian or Asian British–Chinese females, we see that 48 per cent are in continuation group 4. This increases to 79 per cent when restricting to only those who were 18 or younger at the start of their course.

53. We have published interactive webpages for users to explore all of the possible combinations of characteristics and how those groups of students are distributed across the different continuation groups. Alongside this, we have published a downloadable list of all groups of students within each continuation group[17].

Conclusions and further work

54. Analysis and statistical modelling of continuation rates has allowed us to create an ABCS measure of continuation whereby groups of students are assigned to one of four continuation groups based on their modelled continuation rate. Examination of the membership of the continuation groups allows us to understand which groups of students are most at risk of not continuing their course.

[17] See www.officeforstudents.org.uk/data-and-analysis/associations-between-characteristics-of-students/

55. Sensitivity analysis of the model selection method has shown that modelled continuation rates for the larger groups of students are fairly robust to changes – particularly those with very high or very low continuation rates. This means that the resulting continuation groups are also robust to change in the model. Because of this, we are confident that the model is calculating the modelled continuation rates as expected.

56. Further development of this measure would include considering any other characteristics or measures that may also be related to continuation. Given the challenges in creating continuation groups that clearly differentiate between levels of continuation, additional work could be done on the methodology used to create the groups, including testing new ways of defining how big a difference in modelled continuation rates are 'big enough' to be considered as possible start or end points for a continuation group. Additionally, other continuation measure, such as part-time continuation could also be created using the ABCS framework.

We are keen to receive any feedback regarding:

- how the ABCS continuation measure might be used
- the methodology for the statistical modelling
- the methodology for creating the continuation groups.

Please email Annalise Ruck at official.statistics@officeforstudents.org.uk.

Looking at access

57. We have also applied the ABCS framework to the outcome of access to higher education. The access outcome measures the proportion of 18 or 19 year olds entering higher education (sometimes referred to as young participation). Data regarding these students is taken from the DfE's National Pupil Database (NPD) from the summer in which they obtained their key stage four (KS4) qualifications – most commonly, GCSEs. We have then tracked these students through to the start of higher education, where we can determine whether they are in the higher education records two or three years later at the age of 18 or 19. This will capture any level or mode of undergraduate study.

58. We have taken data for pupils who obtained their KS4 qualifications in the summers of 2010, 2011, 2012, 2013 and 2014 (that is, in the academic years 2009-10 to 2013-14) from the NPD. Using KS4 cohorts up to 2013-14 allows us to capture the most recent 19 year old entrants into higher education in the academic year 2017-18. We use KS4 cohorts because they give almost complete coverage of all 16 year olds in England. In addition, combining data from five cohorts allows us to carry out robust analysis, ensuring that there are sufficient students in each of the characteristic groups to allow us to carry out analysis regarding their access behaviour.

59. We have determined five access groups, where groups of students in access group 1 are the least likely to enter higher education, whilst those in access group 5 are the most likely to enter higher education.

Choosing factors to include in the model

60. We have published previous analysis of access rates as part of our access and participation data[18]. This analysis differs from the one used in ABCS as the comparison is made between the make-up of the population of 18 years olds entering higher education and the whole population of 18 year olds, rather than tracking cohorts from the year they obtain their KS4 qualifications. Despite this difference in methodology, the access and participation data gives us an indication that there are differences in access rates between students from different ethnicities, POLAR4 quintiles and IMD quintiles. Therefore, we have carried out further preliminary analysis of the relationship between these characteristics and measures and our own access rates. This has shown all these characteristics are strongly related to access when using tracking of KS4 cohorts.

61. Additionally, we have carried out further analysis of access rates by ethnicity at a more detailed level. This has shown that access rates vary substantially within the Asian or Asian British groups, with access rates being much higher for Chinese and Indian young people than for Bangladeshi and Pakistani young people. Likewise, we found access rates to be substantially higher for black or black British African young people than black or black British Caribbean young people. However, access rates were higher for all groups than for white – English, Welsh, Scottish, Northern Irish or British young people, apart from the mixed – white and black Caribbean, which was slightly lower, and Gypsy, Roma or Traveller, for whom the rate was

[18] See www.officeforstudents.org.uk/data-and-analysis/access-and-participation-data-dashboard/

very low. Due to these substantial differences within broad ethnic groupings, we will include ethnicity in the model at this more detailed level.

62. Examining access rates by sex for young people shows that there is a difference in rates, with females having a higher rate of access to higher education than males.

63. In the same way that we considered including IDACI in the continuation model, we also considered including it for access. As with continuation, we have found that there is a relationship between IDACI and access to higher education. This means that we need to consider the same issues as in continuation about the likely relationships between IMD, IDACI and POLAR4.

64. Finally, we have looked at the relationship between access and free school meal status. This has shown that there is a substantial difference in rates of entry into higher education, with young people who received free school meals having a lower access rate than those who did not. In looking at free school meal status, we have used a variable that flags whether a student received free school meals in the year in which they sat their kS4 examinations (FSM_eligible_spr).

The model

65. As with continuation, we have used a binary logistic regression model to predict the probability of entering higher education. All the variables from the section above have been included. These are: ethnicity, IDACI, IMD, free school meal status (FSM), POLAR4 and sex.

66. Running the model with no interactions has shown that all these variables are statistically significant predictors of entering higher education. Stepwise selection has then been used to determine which two-level interactions should also be included, with an entry and stay criteria of α=0.05. Details of interactions included in the model can be found in Annex C[19].

67. Therefore, the final model is:

$$logit(\pi_i) = \beta_0 + \beta_1 ethnicity_i + \beta_2 IDACI_i + \beta_3 IMD_i + \beta_4 FSM_i + \beta_5 POLAR4_i + \beta_6 sex_i + interactions + \epsilon$$

where i is an individual.

The grouping

68. We have calculated modelled access rates for 6,257 different groups of characteristics. Using the methodology outlined in the 'Creating outcome groups' section, we have created five access groups, where access group 1 are the least likely to enter higher education and access group 5 are the most likely to enter higher education. The groups are as follows[20]:

[19] Available at www.officeforstudents.org.uk/publications/associations-between-characteristics-of-students/

[20] The modelled access rates are only given for groups with 50 or more students. The continuation group might contain rates that are outside the given ranges after smaller groups have been entered.

a. Access group 1 contains groups of students with modelled access rates between 2.3 and 24.4 per cent. This represents 16 per cent of the students included in the modelling.

b. Access group 2 contains groups of students with modelled access rates between 24.6 and 31.8 per cent. This represents 17 per cent of the students included in the modelling.

c. Access group 3 contains groups of students with modelled access rates between 32.0 and 40.0 per cent. This represents 16 per cent of the students included in the modelling.

d. Access group 4 contains groups of students with modelled access rates between 40.2 and 46.1 per cent. This represents 14 per cent of the students included in the modelling.

e. Access group 5 contains groups of students with modelled access rates between 46.3 and 87.6 per cent. This represents 37 per cent of the students included in the modelling.

Independent school pupils

69. There is far less data available about students from independent schools than for pupils from maintained schools. Data about independent school pupils is not collected for the majority of the characteristics and measures we use in the model, causing challenges for how to include them in the analysis. One approach would be to create additional categories for independent school pupils within the characteristics that are not available for them, and include the data that is. However, this would act as a proxy for including an independent school marker. Since we have made the conscious decision not to include school type in the model due to the relationship between school type and prior attainment, we have used a different approach.

70. We have taken the decision not to include independent school pupils in the model, but rather to examine their access rates and consider where these actual rates would place them in the access groups. The only characteristic data we hold for these pupils is sex. Table 1 shows the access rates for all independent school pupils and for male and female independent school pupils. This shows that access rates for all these groups are very high, leading us to conclude that, from the data available, all independent school pupils (regardless of sex) would be assigned to access group 5, and so should be treated the same way as any other groups of students found in access group 5.

Table 1: Access rates for independent school pupils by gender and altogether

	2010	2011	2012	2013	2014
Female pupils	74.3%	74.4%	75.2%	75.2%	75.1%
Male pupils	70.3%	69.6%	69.2%	69.1%	68.8%
All pupils	72.3%	72.0%	72.1%	72.1%	71.9%

71. Details of school types included in the model ('maintained' schools), those classed as independent and those removed from this analysis altogether can be found in Annex A[21].

[21] Available at www.officeforstudents.org.uk/publications/associations-between-characteristics-of-students/

Sensitivity analyses

72. For this analysis, we combine five cohorts of KS4 students. If the access rates were particularly different for a group of students in any of those years, or there is a clear trend in changes over time, we would need to be aware of any impact this was having on our ABCS access measure. Therefore, we have examined access rates across the five cohorts for the characteristics and measures used in the ABCS access model.

73. Across the majority of groups, we see a persistent increase in access rates, showing a steady rise in the proportion of students entering higher education year on year. Since this trend and the relativities of the group rates are so consistent, it does not cause us concern regarding combining the five cohorts. This trend is less apparent when looking at ethnicity, although all groups have seen an overall increase in access rates, this is not always a year on year increase. However, none of these differences are sufficient to cause concern regarding combining the five cohorts.

74. As well as looking at the relationship between continuation and the factors included in the model, it is also necessary to consider the relationships between the factors (i.e. test for multicollinearity). In particular, we had concerns about the relationships between the three area-based measures and between IDACI and free school meal status, since both, in some way, capture deprivation in childhood. To test for multicollinearity, we have looked at correlations between the factors. This showed that there is a strong correlation between IDACI and IMD (ρ=0.887), with no other strong correlations found. When looking at the variance inflation factor (VIF) and tolerance for IDACI and IMD, having run a regression model including all these factors, we found no evidence of multicollinearity between these two factors[22].

75. In developing the statistical model, we have looked at two different selection methods: forward selection and stepwise selection. In each case, we have tried a variety of entry criteria, including α=0.1, α=0.05, α=0.01 and α=0.001. Both selection methods resulted in similar outputs with similar model fit statistics, although the stepwise selection method resulted in fewer two-way interactions being included, and the forward selection method resulted in some interactions being included that were not statistically significant and did not have meaningful estimated coefficients. For this reason, we have chosen to use the stepwise selection method.

76. Having tested various entry and stay criteria, both have been set at α=0.05. For a dataset as large as this, we felt that a value of 0.10 was too liberal and ran a high risk of leading to overfit[23] in the model. Whilst the size of the dataset might usually lead us to conclude that smaller entry and stay criteria would be more appropriate, this is not the case here. This is because we had hoped to include much higher order interaction terms in the model in order to give the model the best chance of robust estimation of the underlying continuation rate but have not been able to due to the very high number of possible higher level interaction terms. We have mitigated this, in part, by selecting more generous entry and stay criteria: α=0.05. Additionally, allowing these less significant terms into the model means they are likely to be

[22] IDACI has a VIF of 4.93 whilst IMD has a VIF of 4.85.

[23] Overfit occurs when a model is too complex and begins to explain random error in the data rather than the relationship between factors.

acting as proxies for some of the higher-level interactions. This also means that we have consistency across the ABCS continuation and ABCS access measures.

Findings from the access groups

77. Having created the access groups, we not only investigate which groups of students fall into which access groups, but also how much information you need to know about a student in order to determine which continuation group they are likely to be in. For example, we find that all students with a Gypsy, Roma or Traveller background will be in access group 1, regardless of which other characteristics they hold.

78. In other cases, we need more information to begin to understand differences in likely continuation outcomes. For example, only 81 per cent of Asian or Asian British - Bangladeshi student are in access group 5, with a further 15 per cent in access group 4 and 4 per cent in access group 3. However, if you look at female Bangladeshi students, rather than all Bangladeshi students, we find that 100 per cent are in access group 5. This tells us that female Bangladeshi young people are more likely to enter higher education than male Bangladeshi young people.

79. When looking at only one characteristic, we note the following:

    a. 43 per cent of males are in access groups 1 and 2, compared to only 23 per cent of females.

    b. 100 per cent of young people of Asian or Asian British–Chinese ethnicity and 98 per cent of young people of Asian or Asian British–Indian ethnicity are in access Group 5.

    c. 86 per cent of all pupils in POLAR4 quintile 5 are also in access group 5.

80. When adding a second characteristic, considering the impact of two characteristic on access to higher education, we note the following:

    a. While only 39 per cent of black or black British Caribbean young people are in access group 5, 73 per cent of female black Caribbean young people are in this group, compared to only 3 per cent of males.

    b. 91 per cent of young people from POLAR4 quintile 1 areas and who received free school meals are in access group 1. This compares to only 48 per cent of all young people from POLAR1 quintile 1 areas.

    c. Whilst only 8 per cent of female young people are in access group 1, 60 per cent of females who received free school meals are in access group 1.

81. We have also looked at how groups of students are distributed across the access groups when looking at three characteristics:

    a. 90 per cent of white (English, Welsh, Scottish, Northern Irish or British) female young people who received free school meals are in access group 1. This compares to only 10 per cent of all English, Welsh, Scottish, Northern Irish or British, female young people.

b. 97 per cent of white, male pupils who received free school meals are in access group 1. This compares to only 29 per cent of white males.

c. 57 per cent of students from IDACI quintile 1 and POLAR4 quintile 1 areas are in access group 1. However when split further by sex, we find that 88 per cent of male pupils from these areas are in access group 1 compared to only 26 per cent of female pupils.

82. The distribution of groups of students for any combination of up to all six of the characteristics and measures included in the statistical model can be explored using our interactive webpages[24]. Data files containing details of which student groups are in each of the access groups are also available on these webpages.

How does this compare to MEM?

83. Our development of the ABCS measure for access is similar to UCAS' multiple equality measure[25] (MEM) but employs a slightly different methodology for some aspects and underlying data which may lead to differences in the analysis. In this section we capture and provide an explanation for these differences.

84. Both ABCS and MEM use a modelling approach which calculates modelled access rates for groups of students with a certain set of characteristics. In both analyses, individual data from the National Pupil Database (NPD) is linked to individual data in higher education in order to determine whether an individual is considered a participant in higher education. For MEM, the NPD data is linked to UCAS data that can tell us whether a student has been accepted to start a course by a provider. For ABCS, a different outcome is examined: the NPD data is linked to the HESA or ILR data which records if a student is present on the course.

85. For these reasons, the definition of access differs between MEM and ABCS: in MEM, a student accesses higher education if they have been accepted by the provider; in ABCS, a student accesses higher education if they have attended the provider. Additionally, ABCS looks at the participation rate for 18 and 19 year olds where MEM only looks at 18 year olds.

86. There are differences in the underlying data used for tracking individuals from school to higher education between the two measures. These are displayed in Table 2. The fact that the years used from the NPD data are different will mean that entirely different individuals, with different characteristics and outcomes, will be used to create MEM and the ABCS access measure. The information below is true of the current version of MEM, although future iterations will consider use of more recent cohorts and will reassess the factors used, looking at other factors that might benefit the model.

[24] See www.officeforstudents.org.uk/data-and-analysis/associations-between-characteristics-of-students/

[25] https://www.ucas.com/data-and-analysis/ucas-undergraduate-releases/ucas-undergraduate-analysis-reports/equality-and-entry-rates-data-explorers

Table 2: Comparison of data used for ABCS access and MEM

	ABCS	MEM
Years of NPD cohorts	2009-10 to 2013-14 (where students sit their GCSEs in the summer of that academic year)	2003-04 to 2007-08 (where students sit their GCSEs in the summer of that academic year)
Age of NPD cohorts	16	16
Location of NPD cohorts	England	England
Higher education data used to match	Data from HESA and ILR	Data from UCAS admissions data
Criteria for participant	Reported in data as attending higher education as an undergraduate (first degree or other undergraduate level)	Reported in data as confirmed and placed as a full time undergraduate
Age in higher education	18 or 19	18
School type	Maintained school pupils only	Separate models for non-independent and independent school pupils

87. The ABCS access measure and MEM both employ a binary logistic regression model. Using the different explanatory variables, the models calculate a modelled access rate for individuals with each possible combination of characteristics. Both models also include two-way interactions between the factors and employ a stepwise selection method with and entry and stay criteria of $\alpha=0.05$ to determine which factors should be included in the model.

88. For both ABCS access measure and MEM, characteristics were selected on a similar basis: that these characteristics should not affect access to higher education, but evidence suggests that these differences exist. All the characteristics chosen have undergone exploratory analysis and have been shown to have a relationship with access to higher education. Table 3 highlights the factors included in the two models, and how many levels there are within those factors. This is particularly important for ethnicity because MEM uses a broad ethnic grouping (Asian, black, mixed, white and other) where ABCS uses a more detailed ethnicity grouping (details in Annex A[26]).

[26] Available at www.officeforstudents.org.uk/publications/associations-between-characteristics-of-students/

Table 3: Factors included in the ABCS access model and the MEM model

ABCS access measure	MEM
Ethnicity (18 levels)	Ethnicity (6 levels)
Free school meals status (2 levels)	Free school meals status (2 levels)
IDACI (5 levels)	-
IMD (5 levels)	IMD (as a continuous variable)
POLAR4 (5 levels)	POLAR3 (5 levels)
Sex (2 levels)	Sex (2 levels)
-	School type (5 levels)

89. For MEM, a slightly different version of the access model is developed using only individual data for independent school pupils because of the lack of data on characteristics for them. This means that there are two access models for MEM: one for state school pupils, which includes all listed explanatory variables, and one for independent school pupils, which has sex as the only factor. For ABCS, we have not included independent school pupils in the model, having calculated access rates for independent school pupils and established that they should be treated the same as those in access group 5, given that their access rates are very high, for both female and male students (see the 'independent school pupils' section for more details).

90. Although the final output created by both MEM and ABCS for access are five groups, where group 1 represents the lowest access rates and group 5 the highest, the method used to construct these groups is very different.

91. MEM groups are quintiles. Quintile 1 represents the fifth of the population least likely to access higher education, and quintile 5 the most likely to do so. ABCS access groups, on the other hand, are not quintiles. Access groups are created in such a way as to maximise the differentiation between the groups. Therefore, the emphasis is more on the group boundaries rather than the number of students represented in each group. The methodology for this can be found in the 'creating outcome groups' section. This results in groups having an unequal share of the population, although the proportions in group 1 to 4 are quite similar:

    - **Access group 1:** 16 per cent
    - **Access group 2:** 17 per cent
    - **Access group 3:** 16 per cent
    - **Access group 4:** 14 per cent
    - **Access group 5:** 37 per cent.

Current conclusions and further work

92. Analysis and statistical modelling of access rates have allowed us to create an ABCS measure of access whereby groups of students are assigned to one of five access groups based on their modelled continuation rate. Looking at the membership of access group 1 allows us to understand which groups of students are least likely to enter higher education.

93. Further development of this measure would include considering any other characteristics or measures that may also be related to access to higher education. Additionally, this framework would allow us to develop measures for different kinds of access, such as for full-time or part-time access only.

We are keen to hear any feedback regarding:

- how the ABCS access measure might be used
- the methodology for the statistical modelling
- the methodology for creating the access groups.

Please email Annalise Ruck at official.statistics@officeforstudents.org.uk.

Conclusions

94. In this analysis, we have created a framework for generating ABCS outcome measures and has used this framework to develop the ABCS continuation measure and the ABCS access measure[27]. In each case, we have employed statistical modelling to calculate modelled rates of the outcome, and used these rates to generate the continuation and access groups.

95. Alongside this report, we have published data downloads that show which groups of students are found in each of the continuation and access groups. Additionally, we have created interactive tools that allow the user to look at the proportion of students holding a set of characteristics in each of the continuation or access groups.

96. It is our intention that providers and other stakeholders test these tools in relation to their effectiveness for identifying groups of students that sit in the lowest continuation and access groups. Potentially, this will allow for the better identification of groups of young people who are least likely to enter higher education, or student groups who are most at risk of not continuing in higher education.

97. At present, these measures are labelled as experimental as this is the first time the methodology has been employed. There is still opportunity for further development of the methodology and to consider other approaches where appropriate. Therefore, we are keen to receive any feedback regarding any of this analysis, so that we might include this is in our thinking around development of this work area.

98. In particular, there is scope for more sensitivity analysis around the number of cohorts used in the analyses, including more rigorous testing of changes in behaviour across the five cohorts. Regarding the grouping methodology, more work could be done on how small groups of students are assigned to outcome groups. Consideration could be given to alternative methods of defining group boundaries for example considering methods that seek to equalise the range of modelled rates covered by each outcome group, rather than the number of people in each group.

99. After we have collected feedback on this experimental methodology and the associated measures, and have had the chance to incorporate any necessary changes, we will finalise the ABCS access and continuation measures. We also intend to develop the ABCS measure for part-time continuation, degree attainment and employment outcomes.

100. Going forward, we intend to use these measures to monitor how the outcomes of groups of students vary. It is our desire to see any gaps in continuation and access closing, and we hope that the ability to specifically target these students will help any interventions reach those students and young people who most need them.

[27] See www.officeforstudents.org.uk/data-and-analysis/associations-between-characteristics-of-students/