Subject choice trends in post-16 education in England

Technical Appendix

August 2024

Michael Scott, Jenna Julius, Sarah Tang and Megan Lucas

The British Academy

NFER

National Foundation for Educational Research

Contents

Notices

This work contains statistical data from ONS which is Crown Copyright. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

The analysis was carried out in the Secure Research Service, part of the Office for National Statistics and draws on data from the National Pupil Database and Individualised Learner Record.

Please note that in order to avoid the disclosure of any personal information, where necessary counts of fewer than 10 students have been excluded and/or rounded figures have been presented to protect the identify of individuals.

1.0 Introduction

This technical appendix accompanies the *Subject choice trends in post-16 education in England* report.

It is structured as follows:

- Section 2.0 provides an overview of how the analysis dataset was created. This includes the subject typology used.

- Section 3.0 covers the approach taken to the regression analysis presented in the report.
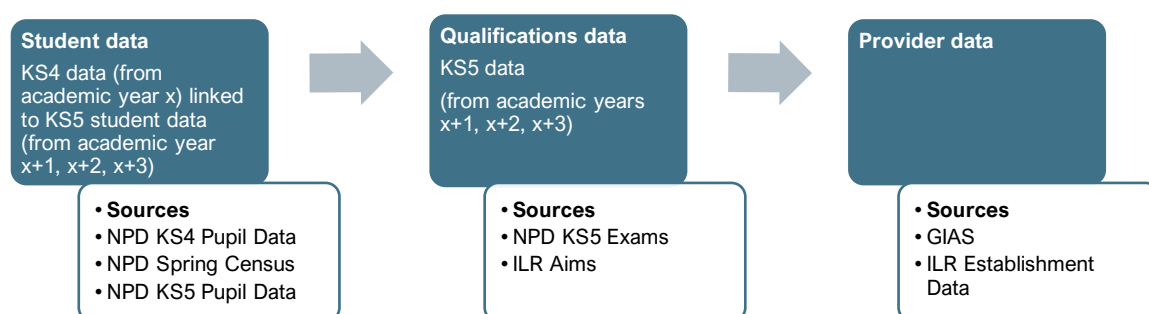
2.0 Data

2.1 Dataset creation

2.1.1 Overview

Our analysis is primarily based on the National Pupil Database (NPD) (Key Stage 4 (KS4) Pupil, KS5 Exams and Spring Census tables) and Individualised Learner Record (ILR) data (Learning Aims table). Additional data was gathered from Department for Education (DfE) Get Information About Schools (GIAS) and a range of post-16 provider administrative datasets to provide details of the education providers where students studied.

We focus on Level 3 qualifications started between the ages of 16 and 18, in the three years following completion of KS4. As such, pupils with AS-Level entries (where the student did not continue to A-level) are included. We include students who are still studying subjects beyond the academic year when they turned 18. In the ILR, we included learning aims which are registered when a student starts pursuing a subject at Level 3, whilst in the NPD, we include any subjects with an exam enrolment. [1] This minimises the extent to which our results are affected by factors that lead to people dropping out or not completing subjects (e.g., difficulty of course) after having chosen the subject.

Figure 1: Analysis dataset creation

Student data
KS4 data (from academic year x) linked to KS5 student data (from academic year x+1, x+2, x+3)

- **Sources**
- NPD KS4 Pupil Data
- NPD Spring Census
- NPD KS5 Pupil Data

Qualifications data
KS5 data
(from academic years x+1, x+2, x+3)

- **Sources**
- NPD KS5 Exams
- ILR Aims

Provider data

- **Sources**
- GIAS
- ILR Establishment Data

[1] We do not include learning aims that are closed within 60 days of being opened.

The spine of the dataset consists of Year 11 students in the NPD who sat KS4 exams in the years 2001/02 to 2019/20. This was then linked to KS5 student and qualifications data, as outlined in Figure 1. As such, we took a cohort-focused approach to structuring the data. We use NPD KS4 Pupil Data to identify groups of students who were likely to start studying for Level 3 qualifications[2] in the year following KS4. Intuitively, we name each cohort for when they would normally be in Year 13 and would have typically sat their final KS5 exams. For example, the '2021/22' cohort will have typically started in Year 13 (or equivalent) in September 2021.

2.1.2   KS5 qualifications data

We combined the KS5 Exams table in the NPD and the ILR Aims dataset to maximise our coverage of Level 3 qualifications. We collated data about the Level 3 qualifications studied in the three academic years that would typically correspond to Year 12, Year 13 and Year 14 for that student to create a single record of the subjects studied. We excluded qualifications linked to apprenticeships but otherwise defined all Level 3 qualifications as being in scope. Any student who did not have a record of studying a Level 3 qualification was excluded.

A- and AS- levels are included from the 2003/04 cohort onwards and other Level 3 qualifications are included from 2007/08 onwards[3]. Data regarding qualifications only studied in 2022/23 (i.e. Year 14) was not available for the 2021/22 cohort at the time of doing the analysis. The trends we present for the 2021/22 cohort are in line with previous cohorts and therefore do not appear to be significantly affected by this.

There are also slight differences in the collection of the ILR aims (which record any aims started by students) and NPD exams (which record exam enrolments), which may impact comparability of vocational and academic qualifications. To maximise comparability, we restrict our sample to ILR record where students who did not withdraw from their qualifications within 60 days.

It is important to note that, for this project, our dataset has been purposefully constructed to recognise the fact that some students will study for multiple qualifications across different academic years. This enables our analysis to holistically consider all the subjects a given student studies at Level 3 even if some records are in different years. It does mean that some students may have studied a given subject in a different year to the one they are assigned in the statistics.

2.1.3   Student characteristics

Student characteristics were identified from the Spring Census in the NPD for when students were in Year 11. This reflects the fact that some characteristics (e.g. free school meals (FSM) status) are not recorded for Level 3 students. Prior attainment is captured from the KS4 Pupil Data.

It is important to note that comprehensive student characteristic data is not available for students who did not complete their Key Stage 4 (KS4) in the state-funded school system. In particular, students who attended independent schools for their KS4 do not have comprehensive characteristic data as this information is not collected from independent schools.

---

[2] Level 3 qualifications are those typically studied for in Years 12 and 13 in England. They include: AS/A-level, advanced apprenticeship, applied general, international Baccalaureate diploma, Level 3 National Vocational Qualification and T Level (Department for Education, no date).
[3] Other Level 3 data is not included for earlier years as subject information was not available.

As outlined in Table 1, it is also important to recognise that over our analysis period, there are notable differences in the measures available for different student characteristics. For example, there have been changes to the identification of Special Educational Needs and Disabilities (SEND). Similarly, there have been a number of breaks in the construction of the series identifying the achievement of a Level 2 at KS4, as shown in the table. These should be borne in mind when interpreting the student characteristic data used in our analysis over time.

Table 1: Student characteristics

Characteristic Variable	Source	Variables*
Gender	Spring Census	Gender (2002-2020)
Special Educational Needs (SEN)	Spring Census	SENStage (2002) SENStatus (2003 – 2005) SENProvision (2006 – 2020)
English as Additional Language (EAL)	Spring Census	Derived from FirstLanguage (2003 – 2006) LanguageGroup (2007 – 2008) LanguageGroupMajor (2009 – 2020)
FSM eligible	Spring Census	FSMeligible (2006 – 2020) FreeSchoolMeals (2002) FSMeligibility (2003 – 2005)
Ethnicity	Spring Census	EthnicGroup (2002 – 2008) EthnicGroupMajor (2009 – 2020)
IDACI quintile	Spring Census	IDACIScore (2002 – 2014) IDACIScore_15 (2015-19) IDACIScore_20 (2020)
Prior attainment	KS4 Pupil Data	Ks4_fiveac (2002 - 2003) - Achieved 5 or more GCSE/GNVQs at grades A*-C Ks4_level2 (2004 - 2013) - Achieved Level 2 threshold (5 or more GCSE and equivalents at grades A*-C) Ks4_level2_ptq (2014) - Achieved Level 2 threshold (5 or more GCSE and equivalents at grades A*-C)(inc S96, WOLF, first Entry and vocational qualification capping) Ks4_level2_ptq_ee (2014-2020) - Achieved Level 2 threshold (5 or more GCSE and equivalents at grades A*-C/9-4)(inc S96, WOLF, first Entry and vocational qualification capping))
KS5 Provider	KS5 Student Data/ ILR	URN, UPIN and UKPRN

Note: The years in this column are the years in which the students in the dataset were in Year 11 as this is the data used for a given student

2.1.4   Education provider characteristics

We sourced provider characteristics (such as region and type) from DfE GIAS data. Education providers are identified for data matching using a mixture of Unique Reference Numbers (URNs) in the NPD KS5 Pupil data and equivalent reference numbers in the ILR aims data. Provider and/or providers characteristic information could not be identified for some students in the data. Where a student was linked to multiple provider reference numbers in the data, we used the one they were most commonly linked to (e.g., had the most qualifications with).

Only providers with at least 20 Level 3 students in any given cohort are counted in the statistics covering providers offering given subjects. This ensured our statistics were not influenced by very small providers which tended to exhibit more variation in subjects offered across years.

2.2   Data considerations

There are some points to note when interpreting the findings which relate to limitations of the data or the construction of the data. The key considerations most pertinent to the analysis are summarised here.

- When extracting this data, some datapoints have been removed to avoid disclosure of personal information. In particular, counts of less than ten are not included. Where a data point is apparently "0", that may well not be the case. The true number could be anywhere between zero and nine. The same is true for figures like "0%".

- As a means of structuring our dataset we considered all exams data associated with a student in the three years following their Year 11. Any exams taken in this period were assigned to a cohort year which would be Year 13. Because of this, our dataset is not comparable to other published statistics (e.g., Department for Education A-level and Level 3 statistics, Joint Council for Qualifications exam entries) which are mostly constructed based on the qualifications studied in a given academic year.

- Where students have missing data on their characteristics (or provider characteristics), they have been excluded from analysis where that data was needed. (e.g. If Gender is unknown, that student is not included in the calculation of the proportion of students of given subject who are female or male.)

- In the descriptive data students may be excluded in the percentages for one characteristic (where their data is missing) and not on another. This is why some counts may not match between statistical tables.

- In the regression analysis, students were dropped from all models if any of the data on their characteristics is missing, because that data is needed for all models.

- Students from independent schools are excluded from all analysis which uses pupil characteristic information, both descriptive statistics and regressions. They are excluded in statistics on subjects studies by students and provider statistics too.

- The KS5 Pupil Data in 2015/16 and 2018/19 was impacted by high numbers of missing provider reference numbers. Where possible provider reference data was gathered from the Spring Census or other years of data.

2.3 Typology of subjects

A very wide set of subjects have been studied at Level 3 over the last 20 years in England.

For this project, it was important to build a typology of subjects which:
1. Collated smaller subjects and courses into groups.
2. Established a definition for which subjects are Arts, Humanities, Social Science, and STEM subjects.

The typology used for this study is outlined in Table 2.

Table 2: Typology of subjects

Major Groups	A-levels, including Minor Groups	Vocational Qualifications Mainly based on Sector Subject Area (SSA) Tier 1 and 2 Codes[4]
Humanities **(13 A Level + 2 VQ)**	Single Subjects: 1. History 2. English Literature 3. English Language 4. English Literature and English Language 5. Religious Studies 6. French 7. Spanish 8. German Minor Groups: 9. Other Languages (European) 10. Other Languages (Non-European) 11. Classical Subjects (Languages) 12. Classical Subjects (Non-Languages) 13. Other Humanities Subjects	1. History, Philosophy and Theology VQs • SSA 10 - History, Philosophy and Theology 2. Languages, Literature and Culture VQs • SSA 12 - Languages, Literature and Culture
Social Sciences **(8 + 5)**	Single Subjects: 1. Sociology 2. Geography 3. Economics 4. Law 5. Government and Politics Minor Groups: 6. Psychology 7. Business Studies (including 'with Economics' variants) 8. Other Social Science Subjects	1. Core Social Science Subject VQs • SSA 11 - Social Sciences 2. Education and Training VQs • SSA 13 - Education and Training 3. Business, Administration and Law VQs • SSA 15 - Business, Administration and Law 4. Psychology VQs • Any qualification with 'Psychology' in title. 5. Other Social Science VQs • SSA 03.4 Environmental conservation

[4] With qualifications classified based on SSA tiers, there is a small minority of qualifications (mainly before 2010) where there is only a 'Tier 1' SSA code and no 'Tier 2' code. For example, a qualification may have a 'Tier 2' code of '01', which relates to 'Health, Public Services and Care', but the data does not distinguish whether the code is 01.1, 01.2, 01.3, 01.4 or 01.5. For the vocational qualification where the SSA Tier 1 code is split in our scheme, we have:
- Treated SSA 03 and SSA 04 as "Agriculture, Horticulture and Animal Care VQs" and "Engineering and Manufacturing Technologies VQs"
- Treated SSA 08 as "Other VQs"
- Treated SSA 14 as "Other VQs"
In the case of the latter two groups, we are operating on the principle that the data does not positively identify them as social science subjects.

		• SSA 08.2 Travel and tourism • SSA 14.2 Preparation for work • Any qualification with 'Digital Marketing' or 'Employment Related Services' in the title.
Arts **(5 + 1)**	Minor Groups: 1. Media/Studies/Film TV (incl Film Studies) 2. Art and Design Subject (subjects with 'Art and Design' in title) 3. Performing Arts 4. Music (incl Music Technology) 5. Design and Technology Subject 6. Other Arts	1. Arts, Media and Publishing VQs • SSA 09 – Arts, Media and Publishing
Science, Technology, Engineering and Maths (STEM) **(6 + 3)**	Single Subjects: 1. Biology 2. Chemistry 3. Physics Minor Groups: 4. Maths (all types of Maths A Level and Further Maths) 5. IT and/or Computer Studies 6. Other STEM Subjects	1. Science and Mathematics VQs • SSA 02 - Science and Mathematics 2. Engineering and Manufacturing Technologies VQs • SSA 04 - Engineering and Manufacturing Technologies 3. Information and Communication Technology VQs • SSA 06 - Information and Communication Technology
Other subjects **(2 + 5)**	Single Subjects: 1. Physical Education 2. General Studies	1. Health, Public Services and Care VQs • SSA 01 - Health, Public Services and Care 2. Agriculture, Horticulture and Animal Care VQs • SSA 03 - Agriculture, Horticulture and Animal Care (Except SSA 03.4, Environmental conservation, which we treat as Social Science). 3. Construction, Planning and the Built Environment VQs • SSA 05 - Construction, Planning and the Built Environment 4. Retail and Commercial Enterprise VQs • SSA 07 - Retail and Commercial Enterprise 5. Sports and Foundations (VQs) • SSA 08.1 Sport, leisure and recreation • SSA 14.1 Foundations for learning and life
Totals	**5 major subject groups and 51 minor subjects/subject groups.** **(35 AS/A-level subjects or minor subject groups, 16 VQ subject groups).**	

Table 3 defines which AS/A-Level subjects make up the minor groups where they have been listed above.

Table 3: Minor subject groups

Minor Group (Same order as in table above)	AS/A-level Subjects (Called 'Mapping' in NPD)
Other Languages (European)	Italian, Polish, Russian, Turkish, Portuguese, Modern Greek, Dutch, Irish, Welsh
Other Languages (Non-European)	Chinese, Arabic, Urdu, Japanese, Punjabi, Persian, Bengali, Gujarati, Hindi
Classical Subjects (Languages)	Latin, Classical Greek, Other Classical Languages
Classical Subjects (Non-Languages)	Ancient History, Classical Civilisation, Classics
Other Humanities Subjects	Critical Thinking, History of Art, Logic/Philosophy, Archaeology, European Studies, Communication Studies
Psychology	Psychology, Psychology (as science)
Business Studies (including 'with Economics' variants)	Business Studies, Business Studies & Economics
Other Social Science Subjects	Accounting & Finance, World Development, Home Economics, Anthropology, Social Science: Citizenship
Media Studies	Media/Film/TV Studies, Film Studies
Art and Design Subject (subjects with 'Art and Design' in title)	Subjects with 'Art and Design' prefix and subjects with the same name: Fine Art, Photography, Graphics, Textiles (Arts), 3D Studies, Critical Studies, Art
Performing Arts	Drama & Theatre Studies, Dance, Expressive & Performance Arts, Theatre Studies
Music (incl Music Technology)	Music, Music Technology
Design and Technology Subjects	Subjects with 'Design and Technology Prefix': Product Design, Food, Textiles, Systems and Controls, DT, Design, Communication
Other Art Subjects	Creative Writing, Textile/Fashion
Maths (all types of Maths A-level and Further Maths)	Maths, Further Maths All maths module titles: Additional Maths, Use of Maths, Statistics, Pure Maths, Mechanics, Decision Maths
IT and/or Computer Studies	IT, Computer Studies
Other STEM Subjects	Geology, Environment Science, Electronics, Science, Science for Public Understanding, Science in Society

3.0 Methodology

3.1 Approach

The regression analysis investigates the relationships between different characteristics (student and provider), time-fixed effects and subject choices at A/AS Level. Specifically, we use regression modelling to estimate the extent to which different factors are associated with the likelihood of a student studying each subject, whilst controlling for differences in how other characteristics that may also drive differences in subject choice.

Due to the categorical nature of our dependent variable (whether a student chooses a subject or not), a logistic regression model was used rather than a linear model. The regression was specified as follows, with a given subject or subject group as our dependent variable of interest and dependant variables summarised in Table 4:

$$P(Studying\ Subject) =$$
$$F(\ B0\ +\ B1\ \times\ Student\ Level\ Factors\ +\ B2\ \times\ Provider\ Level\ Factors$$
$$+\ B3\ \times\ Geographic\ Factors\ +\ B4\ \times\ Binary\ Variables\ for\ Each\ Year) + e$$

P() means "the probability of", F(x) is the logistic function and e is a residual term.

Regressions were estimated using a generalised maximum likelihood approach. Complete case analysis was used, whereby only students with non-missing data were included. This was because missingness tended to be correlated across different variables. Standard errors are clustered for students in the same provider.

Table 4: Independent variables in model

Variable group	Variables
Student characteristics	Gender, Ethnic Group, English as Additional Language, Free School Meal Status, Special Educational Needs Status, Aggregated Prior Attainment at GCSE, IDACI Quintile
KS5 provider variables	Provider Type, Provider Size
Geographical variables (from provider data)	Region, Rural/Urban
Time	Binary variable for each year except the first year.

Regressions were estimated separately for:
- All major subject groups except 'Other' (for both AS/A-level only and all Level 3 qualifications)
- For the top five humanities and top five social science subjects[5] (based on the total number of students studying these subjects between 2003/04 and 2021/22).
- A selected sub-set of other individual subjects or minor groups, where there had been key policy changes over our analysis period.

---

[5] All English A-levels were counted as one subject

Due to computational limitations, regressions could not be estimated across all subjects covered in the descriptive analysis.

3.2   Sample

For the majority of our regression analysis, we use a dataset consisting of students studying AS/A-levels between 2003/04 and 2021/22, covering some 5.4 million students. We removed those studying KS5 at independent schools or providers, reducing our sample to 4.8 million students (89% of our original AS/A-level dataset). Other students that did not have complete data (for the independent variables) were also removed from the analysis dataset resulting in 4.5 million students (82% of our original AS/A-level dataset).

5.7 million students were included in the regressions looking at all Level 3. This covered the cohorts between 2007/08 and 2021/22 and students were excluded from the dataset on the same basis as outlined above.

3.3   Sensitivity analysis

We tested the robustness of our findings using a number of sensitivity analyses and found that our results were largely comparable across the different specifications analysed. This included estimating models separately by gender and for different sub-sample (e.g., different approaches to handling missing data).

References

Department for Education (no date) *What qualification levels mean*, *GOV.UK*. Available at: https://www.gov.uk/what-different-qualification-levels-mean/list-of-qualification-levels (Accessed: 4 March 2024).