Socio-economic status and subject choice at 14: do they interact to affect university access*

Jake Anders**, Morag Henderson**, Vanessa Moulton**, and Alice Sullivan**

**UCL Institute of Education

29 November 2017

Acknowledgements

This report details work conducted as part of Nuffield Foundation grant EDU/42169 (Principal Investigator: Jake Anders). Parts of the work under this grant were carried out in collaboration with Morag Henderson, Alice Sullivan and Vanessa Moulton, working on research funded by the Economic and Social Research Council. In particular, the analysis reported in Section 4 was led by Morag Henderson. The Nuffield Foundation is an endowed charitable trust that aims to improve social wellbeing in the widest sense. It funds research and innovation in education and social policy and also works to build capacity in education, science and social science research. The Nuffield Foundation has funded this project, but the views expressed are those of the authors and not necessarily those of the Foundation. More information is available at www.nuffieldfoundation.org.

Nuffield Foundation

*Contact details: Jake Anders (jake.anders@ucl.ac.uk), UCL Institute of Education, 20 Bedford Way, LONDON WC1H 0AL.

Executive Summary

There is a large socio-economic status gap in higher education (HE) participation in England. However, most evidence suggests that this is driven by inequality that emerges before the point of application. It has been suggested that one such source of inequality is the subjects and qualifications studied by young people while still at school. The importance of this factor for young people's chances of progressing to HE in general, and to highly selective HE institutions in particular, has increasingly attracted the attention of policy-makers. This has been most notable in the UK Government's introduction of the English Baccalaureate performance measure for schools at age 16, and the introduction of performance in Russell Group "facilitating subjects" at A-Level for schools at age 18. However, this area is under-studied in the academic literature.

This project aimed to address this gap using a combination of survey and administrative data on a recent cohort of English students. It analysed the subject choices taken by young people at age 14 (affecting subjects and qualifications studied for examinations predominantly at age 16) using statistical analysis to estimate the subsequent importance of subject choice in the probability of attending university or a highly competitive university. It also considers the association between socio-economic status and young people's subject choices, and the extent to which this acts as a transmission mechanism between socio-economic status and inequality in attendance at university.

Overall, the findings of this project highlight a number of important implications for schools and policy-makers:

- There are substantial socioeconomic differences in the subjects that young people study from age 14 to 16. Young people from advantaged households take more selective subjects, have higher odds of doing three or more facilitating subjects, higher odds of studying a full set of EBacc-eligible subjects (including English, Maths, History or Geography, two sciences and a modern or ancient language), but lower odds of taking Applied GCSEs (e.g. Applied Hospitality, Applied Health or Applied Manufacturing) than less advantaged young people. These differences do not simply reflect prior academic attainment but persist even once this has been held constant. These persistent differences confirm the potential for curricular differences at 14 to exacerbate inequalities rather than simply reflecting existing inequalities.
- These differences are partly associated with the schools in which young people find themselves at this point, not just their individual characteristics. We found that there were important differences by school characteristics, which may be a result of differential opportunities, subjects offered and within school policies. As such, we should be sceptical of considering young people's subjects of study purely in terms of 'choice' (Woods, 1976). They are, at most, constrained

choices, potentially both for individuals and for schools (see next point).

- Even holding other factors constant, pupils in non-selective schools within selective local authorities study a less academically selective set of subjects. This suggests that expanding selective education will both increase inequality in, and decrease the average level of, academic selectivity of subjects that young people study.
- When we consider university entry, and admission to high-status universities in particular, there are large raw differences associated with studying more academically combinations of subjects. However, once differences in young people's backgrounds and prior attainment associated with these differences in subjects studied are taken into account, these differences are, at most, small. A large policy focus on incentivising schools to provide and individuals to study particular combinations of subjects is unlikely to have more than a marginal difference.
- The results for studying the full set of EBacc subjects and for studying any applied subjects do show residual associations with university attendance, suggesting the view that they may have particular importance is not without merit, a finding that concords with other research focussing on subject choice at a later point in individuals' educational careers (Dilnot, 2016). Nevertheless, it is important to emphasise that the differences are still not large, suggesting that the weight that has been placed on the EBacc by policymakers has been exaggerated.
- Our findings suggest that if young people from different socioeconomic backgrounds were studying a more similar curriculum between ages 14 and 16 it would be unlikely to make much of difference to the inequality in university entry highlighted by previous studies. This does not mean that ensuring pupils have the same opportunities to choose their curriculum post-14 regardless of their background is not important. We may regard this as important in itself for reducing socioeconomic differences in earlier educational trajectories, and also having the potential to make a difference to inequality in university going at the margin. However, we certainly should not regard reforms in this space as any kind of 'silver bullet'.

Contents

1 Introduction

Young people's subject choices at age 14 may have important consequences for future academic and labour market outcomes, since they in turn affect the qualifications to which they can easily continue in post-compulsory education. Choosing the 'wrong' set of options at this point may have long term consequences (Iannelli, 2013). This is a particularly important issue in an English context, where specialisation of the curriculum occurs earlier than in many other countries (Hodgson and Spours, 2008).

The choices that individuals face seem likely to be shaped by the schools in which they find themselves at this point in time, just as previous work has found that pupils' options are restricted depending on where in the country they live (Open Public Services Network, 2015). Schools may not offer certain subjects (Jin et al., 2011) and often guide their pupils towards certain paths (McCrone et al., 2005), for example requiring that a wider set of core subjects be studied, or preventing pupils from taking certain combinations of options. This implies that schools potentially have an important influence in this regard. However, schools do not set such requirements in isolation. They face significant constraints most obviously from government policy but also in responding to what they can offer given the make-up of their student body. For example, they cannot viably offer an optional subject that only a handful of pupils wish to study. Similarly, the local education market, especially the presence of selective schools, may influence other schools' behaviour.

A major part of the 2010-2015 UK government's education reforms in England was a focus on the curriculum that pupils study from ages 14-16. Most high profile was the introduction of the English Baccalaureate (EBacc) performance measure for schools. Since schools were now judged on the proportion of pupils getting a "good pass" in the subjects that made up this measure, it incentivised schools to encourage pupils to study this set of "subjects the Russell Group identifies as key for university study" (Gibb, 2011). Young people's parents also see the choices their children are making at this point in time as important, with 93% of the parents of the Next Steps (formerly known as the Longitudinal Study of Young People in England) participants saying they see subject choices at age 14 as "very important" or "fairly important" for the educational options their offspring will have open to them subsequently. However, there does not appear to be good quantitative evidence about the importance of studying a complete set of subjects, per se. Indeed, concern has been expressed in some quarters that a particular focus on a set of subjects such as the English Baccalaureate (EBacc) might 'crowd out' other subject combinations, such as a full set of separate sciences, that are also potentially important for individuals' future educational opportunities.

Over the past twenty years governmental efforts to promote social mobility have included widening access to higher education as a major focus. This is in an attempt to give more individuals the opportunity to benefit from the economic returns to a university degree (Walker and Zhu, 2011). Despite this, there remains a significant level of socioeconomic inequality in access to universities. Previous analyses using multiple sources of data have established that much of this gap in enrolment is explained by prior academic attainment (Chowdry et al., 2013) and by differences in application behaviour (Anders, 2012a) but that there remain some differences, particularly in access to highly competitive universities (Boliver, 2013).

This report proceeds as follows. Section 2 reviews the previous literature on factors that shape young people's subject choices and the possible consequences flowing from this. Section 3 outlines the data used in the course of this project provides basic description of the patterns of subject choice. We build on this in Section 4 by jointly considering individual predictors of the subjects that young people study; Section 5 builds on this to also consider the importance of schools in shaping young people's subject choices at this point in time. From these findings regarding differences in the subjects that young people study, the next two sections consider the potential consequences of these differences. Section 6 employing a matching approach to consider the importance of studying particular combinations of subjects for young people's chances of entering university. Section 7 then builds on this by exploring the extent to which the subjects individuals study between ages 14 and 16 explain socioeconomic inequality in university attendance. Finally, Section 8 draws out overarching conclusions from this project.

2 Background

By and large, England has a system of within, rather than between, school curricula differences. This is associated with smaller socioeconomic differences in the curricula that individuals take (Chmielewski, 2014). Nevertheless, there are significant differences in the subjects that young people study depending upon their background. This project has considered subject choices at this point in time because, in the English context, it is the first time that individuals get to express a preference for the subjects they study. It is also a point at which all young people are still in compulsory education for two more years. Unlike studying post-16 subject choices, there remains something of a common core to the curriculum, allowing a focus on how choices about non-compulsory subjects seem to affect future plans.

Previous studies that have explored the determinants of subjects studied between ages 14 and 16, have tended to highlight that three important characteristics in explaining subject choices at this age are gender (Bell, 2001; Francis, 2000; Jin et al., 2011; Sullivan et al., 2010), prior attainment (Davies et al., 2008; Jin et al., 2011) and socioeconomic background (Davies et al., 2008; Jin et al., 2011). Bell (2001) considered changes in the uptake of combinations of age 14-16 subjects by gender and prior attainment and how this changed with the introduction of the National Curriculum. Davies et al. (2008) used the 1998 Year 11 Information System (Ye11IS) data to examine the probability of taking GCSEs in optional subjects (specifically: Business Studies, French, Geography, German, History and Home Economics), finding that 'ability' has the strongest influence on subject choice but for some subjects social class exerts more of an effect than gender. Using a more recent cohort, Jin et al. (2011) find that girls are more likely to study modern foreign language at school and less likely to study all three sciences separately; these associations remain after taking into account prior attainment. Furthermore, those with more educated parents are more likely to study triple science and to stay on in full-time education after Year 11, however these effects are not significant after controlling for prior attainment. Sullivan et al. (2010) also make use of Next Steps data to examine the social structure of the Key Stage Four curriculum in England. They examine how the subjects that young people like or dislike shape the choices that they make and how they are influenced by those around them. In addition, they note a gender difference in vocational subjects and differences by ethnicity for triple science and religious study participation.

Previous work on the importance of subject choice during secondary school has also focussed on specific elements of the decision, for example considering whether young people study Science, Technology, Engineering and Maths (STEM)-related subjects (Tripney et al., 2010; Codiroli, 2015). In the case of STEM, this reflects a concern that there is a gender gap in uptake of such subjects, although Codiroli (2015) highlights that this not be the case among individuals from advantaged backgrounds.

Aspects of the role that schools play in shaping subject choice have also been considered. Jin et al. (2011), using Next Steps, document significant variation in the kinds of qualifications offered by different schools. In addition, they note that “19% of pupils were unable to take subjects they would like to study at Key Stage 4” (Jin et al., 2011, p.63), with the most common reason being that their school did not offer the subject (just over 30% of such cases). They identified large differences by school, where some schools offer courses in both academic and vocational choices in Year 10 while others only offer academic courses. Davies et al. (2008) also considered the influence of school context, noting associations between school cohorts and probability of taking subjects, for example the proportion of children who are eligible for free school meals in the school has an effect on the probability of taking certain subjects.

There is a small amount of international evidence on the effects of the sex composition of co-educational schools and classes. Hoxby (2000) uses data on schools in Texas to show that a higher proportion of boys in the class depresses the attainment of both male and female students in both maths and English. Hoxby suggests various possible mechanisms for such peer effects, including classroom disruption and changes in classroom atmosphere. Israeli research also suggests that a high proportion of boys in a year group is linked to worse academic outcomes for both girls and boys (Lavy and Schlosser, 2011). Van Houtte (2004) produces similar findings for Belgium. Proud (2014) uses PLASC/NPD data for England and finds that a higher proportion of girls in the class has a negative effect on boys attainment in English, while a higher proportion of girls has a positive effect on both girls and boys science attainment. Sullivan (2009) found that teenage girls in the 1970s rated their abilities in maths and sciences higher if they went to an all-girls school. Boys on the other hand rated their abilities in English higher if they went to an all-boys school (Sullivan, 2009). Similarly, boys and girls who attended single-sex schools showed increased attainment in gender-atypical subject areas Sullivan et al. (2009), suggesting that single-sex schools may contribute to breaking down gender stereotypes.

It is well established that students attending schools with a high proportion of peers of low social status or low academic ability are at a disadvantage (Coleman et al., 1966; Henderson et al., 1978; Mortimore, 1988; Rutter, 1982; Smith et al., 1989; Willms, 1986). Recent research has suggested that school-SES has no direct effect on individual level attainment except via the academic composition of the school (Marks, 2015). The mechanisms behind school composition effects have not been established empirically. School composition effects may reflect peer group processes (for example if lower-attaining peers are more disruptive). School composition may also influence teachers and the curriculum, as teachers seek to provide a curriculum and pedagogical style which they deem appropriate for the population of the school as a whole.

The importance of the subjects young people study while at school for their chances of progressing to Higher Education (HE), in general, and highly selective HE institutions, in particular, has increasingly attracted the attention of policymakers (Gibb, 2011). The policy attention stems from a concern that young people are making subject choice decisions (or being channelled towards decisions) that are reducing the probability of participating in Higher Education and that this is more likely to be the case for those from less advantaged backgrounds. Indeed, previous work has suggested that when high achieving young people from less advantaged backgrounds are provided with more information on how best to prepare for university applications their decisions improve (Borghans et al., 2013; Hoxby and Turner, 2013). Although these previous studies cited did not specifically cover advice about subject choice, there is a similar logic of improving educational decisions in our setting. Indeed, previous research has highlighted that choosing the 'wrong' curriculum at this point may have long term consequences in terms of occupational status acquisition (Iannelli, 2013); educational progression seems one plausible mechanism for this. In a different context, evidence from Belgium suggests that subject choice has an influence on the gender gap in the labour market (Duquet et al., 2010).

In their analysis of administrative data (National Pupil Database records linked to data from the Higher Education Statistics Authority), Chowdry et al. (2013) find large socioeconomic inequality in university attendance. They find that that individuals of the top fifth of their sample are more than 40 percentage points (% pts.) more likely to start university than those in the least advantaged fifth, with larger gaps towards the upper end of the distribution. They also find differences of more than 30% pts. in the probability of attending a high status university between the same groups. They go on to find that much of this difference is explained by attainment that emerges earlier in the education system. However, while this use of administrative datasets has clear advantages, the relative weaknesses of socioeconomic measures available may mean this analysis understates inequality on this basis. Other work using survey data with richer information on socioeconomic status has found that some socioeconomic inequality in the probability of attending university remains even after prior attainment is controlled for, although much may be explained by differences in application behaviour (Anders, 2012a). Work by Boliver (2013) suggests that some differences in university entry may remain, although this analysis of Universities and Colleges Admissions Service (UCAS) data is only able to control for prior attainment at age 18, not earlier in the education system. Taken together, there is evidence of residual inequality in university entry (and entry to high-status institutions) although the point during young people's educational careers where this emerges is not clear. As such, it seems relevant to consider other potential mechanisms through which this might be happening.

3 Data

This project has used data from both Next Steps (a representative longitudinal study formerly known the Longitudinal Study of Young People in England) and the National Pupil Database (NPD; an administrative dataset owned by the UK's Department for Education).

Where using the NPD, we focus on the sample from mainstream English state-funded schools for the academic year 2005-06. This includes comprehensive academic attainment data from national examinations in England. Rather than self-reports about subjects of study we use the observed information about which GCSEs (or equivalents) young people have entered at age 16. The advantages of administrative data are clear, in that we know our information about full cohorts within schools. It also includes some basic data providing a proxy for young people's socioeconomic background, namely whether they are eligible for free school meals (FSM) and the deprivation status of their neighbourhood (Chowdry et al., 2013).

However, this is obviously less fine grained than background characteristics available in survey data. Where possible, we test the robustness of our results to this limitation by replicating the analysis using Next Steps (discussed below). While the results do not replicate exactly (especially regarding school-level relationships), this is perhaps unsurprising given differences in the measurement instruments and the fact that only a small number of students within each school are surveyed; furthermore, many of the same broad patterns are evident. Where this is not the case, differences are noted and our confidence in these findings is reduced.

Next Steps follows a cohort of young people born in 1989-90 from age 14 through to age 20. The survey has a clustered design based around schools, so that young people are randomly selected for inclusion within randomly selected schools (albeit with some oversampling). It includes annual interviews throughout with the young people themselves, interviews with their parents (for the first four years), and linked administrative data about young people's academic attainment (from the National Pupil Database, discussed above). Using the responses from the parental questionnaires provides high quality data on young people's socioeconomic background, based on questions about family income, parental education, and occupational status.

Importantly for this work, it also includes self-reported information on subjects that young people are studying at age 14 (academic year 2004/05-2005/06). We use these to generate the subject choice classifications that we use as 'treatment' variables, of which we attempt to assess the intrinsic importance for university outcomes. To provide context regarding the subjects that young people in this cohort studied, Figure 1 shows the proportion of young people selecting particular GCSE subjects. The most popular GCSE selected (recalling that English and Maths are excluded from this

Table 1: Average academic performance at age 14 of pupils studying each GCSE subject ranked in ascending order - NPD Data

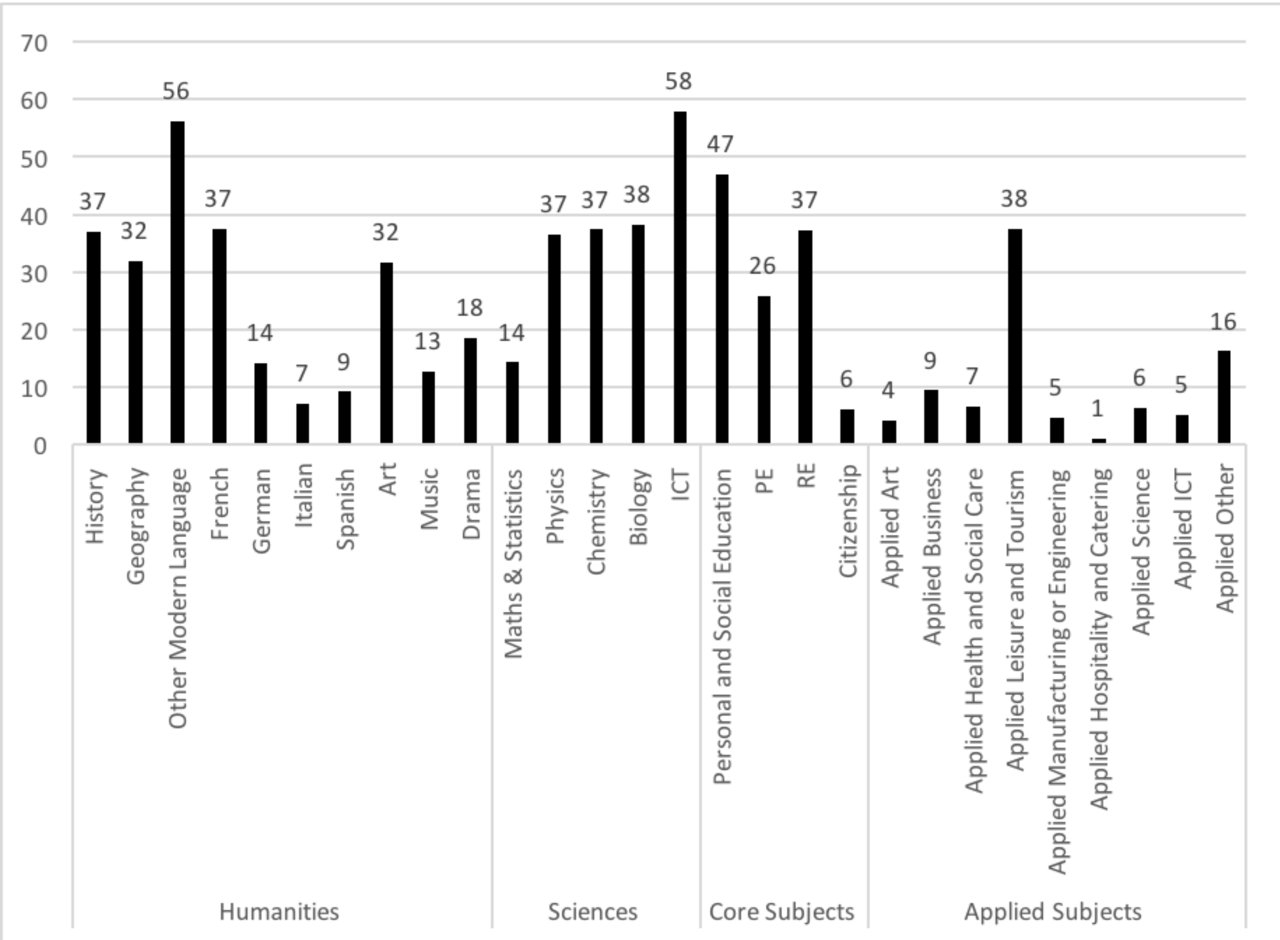
Subject	Academic Selectivity Score
Single Science	31.14
Applied Science	31.77
Applied Home Economics	32.99
Other Foreign Lang.	33.85
Applied Business	34.46
Applied Media	34.78
Applied Office	34.82
Other Applied	34.82
Citizenship	35.29
Maths	35.32
English	35.35
Art	35.39
Design Technology	35.48
Drama	35.57
Applied IT	35.58
Double Science	35.64
Religious Education	35.92
Information Technology	36.21
Geography	36.63
History	37.17
French	37.38
Music	37.65
Spanish	37.71
German	38.17
Italian	38.47
Biology	41.40
Chemistry	41.70
Physics	41.75

Notes: Constructed by calculating average point scores in KS3 tests in English, maths and science at age 14 among all individuals who study each subject.

analysis since they are compulsory) is ICT (58%), followed by Modern Foreign Languages (56%). The least popular subject is Applied Hospitality and Catering (1%).

In both of the datasets we develop an overall, continuous measure of the academic selectivity of the subjects that a pupil studies from age 14-16, based on the prior academic performance of the pupils that choose to study each subject. We assign each subject the average score in Key Stage 3 (KS3) compulsory tests at age 14 of those pupils that report they are studying that subject. KS3 tests are taken roughly contemporaneously with subject choice decisions, so they seem the most appropriate measure to use in this way. The score for a range of subjects is reported in Table 1 using NPD data and in Table 2 for Next Steps. We see that those with the highest levels of KS3 attainment are more likely to study subjects such as languages, while those with lower levels are more likely to take applied subjects of various types.

Figure 1: Proportion of pupils who report studying each GCSE subject



Notes: Adjusted using LSYPE-provided survey design and attrition weights. Sample: Wave 2 respondents with non-missing data on subject choice variables.

Table 2: Average academic performance at age 14 of pupils studying each GCSE subject ranked in ascending order - Next Steps Data

Subject	Academic Selectivity Score
Applied Hospitality	29.28
Applied Leisure	29.91
Applied Health	30.45
Applied Manufacturing	30.55
Applied Art	31.36
Other Applied	32.12
Applied Science	32.62
Other Foreign Lang.	33.31
Applied Business	33.31
Citizenship	33.48
Art	33.62
Drama	33.67
Physical Education	33.78
English	33.87
Maths	33.87
Information Technology	34.15
Personal Social Health Ed.	34.55
Religious Education	34.67
Applied IT	34.75
Music	34.95
Geography	35.23
History	35.66
Biology	35.67
Chemistry	35.69
Physics	35.82
French	36.27
Spanish	36.31
Italian	36.65
Statistics	36.87
German	37.46

Notes: Constructed by calculating average point scores in KS3 tests in English, maths and science at age 14 among all individuals who study each subject.

We next convert this into an individual-level, rather than a subject-level, measure. To do so, we sum the top eight most academically selective subjects that each individual studies. A maximum of eight subjects are used to create this measure in order to stop individuals taking a large number of low-selectivity subjects ending up with a high selectivity score. This follows the logic used in the construction of 'capped' GCSE points scores. Thus, individuals who take a combination of academically selective subjects end up with a high score, while individuals who take a combination of less selective subjects are assigned a low score. For ease of interpretation, we standardise this score among the sample used in this paper, so that it has a mean of zero and a standard deviation of one.

We acknowledge the slight circularity in the calculation of this measure for some of our modelling – highly academically selective subjects are those studied by those with high prior attainment – however, we still think the the extent to which other factors explain part of the this variation means this is still an interesting exercise. In addition, we explore the role of schools in explaining the variation in young people making specific sets of decisions that have previously been argued to be important for future outcomes.

The Russell Group (a self-selecting group of highly selective UK universities) produces an annual document called "Informed Choices" providing advice on the kind of subject choices that will give young people "the most options" when it comes to accessing such universities (Russell Group, 2013). It mainly focuses on advice regarding post-16 subjects, highlighting the importance of the 'facilitating' subjects of English Literature, Maths, History, Geography, Languages, Physics, Chemistry and Biology. Our focus here is on subject choices at age 14-16, for which the Russell Group provides much more general advice. Instead, we focus on whether young people are studying at least three of what might be considered pre-cursors to these post-16 'facilitating' subjects: English, Maths, History, Geography, Languages, Physics, Chemistry and Biology.

As something of a comparison, we consider whether individuals studied for any applied GCSEs. These were introduced in the 2002 Education Act, as part of a policy to increase the diversity of the 14-19 curriculum. However, this policy has since been criticised, with some of these qualifications having their equivalence to GCSEs in performance tables downgraded since this period. Those who did so tend to be less advantaged and have lower prior attainment than those who did not. According to Table 2, they are certainly among the least academically selective subjects, which is hardly surprising as they were used as alternatives to more academically focused subjects for individuals for whom these seemed less appropriate.

We consider the importance of studying the full set of subjects required to be eligible for the English Baccalaureate (EBacc). Recent government policy has sought to incentivise more young people to take this combination of subjects arguing that this provides young people with the skills they need for the future. Others have countered that this approach may be harmful because the focus on ensuring passes in these subjects may be to the detriment of other subjects. For a pupil to count towards their school's EBacc measure they must achieve a C grade or above (often referred to as a 'good pass') in the following GCSE subjects: English, Mathematics, History or Geography, two sciences and a Modern or Ancient Language. However, the introduction of this performance measure comes after the cohort we consider took their GCSEs. This strengthens our approach since it eliminates the possibility that individuals took these subjects specifically in order to achieve the EBacc, which may increase any selection issue; constructing an indicator of studying EBacc subjects artificially for this cohort should give us a cleaner estimate of whether studying the required subjects improves university entry chances in and of itself. We construct a binary measure according to whether pupils study the full set of subjects that would make them eligible for the EBacc if they a) go on achieve at least a grade C in all of them and b) were in a later cohort when the measure had been introduced. We find that one third of the sample studied subjects that would have make them eligible for the EBacc in later years.

We also consider whether individuals study specific elements that make up the EBacc. We assess whether individuals study two or more sciences i.e. two of Physics, Chemistry and Biology as separate subjects or a combined 'double' award in sciences during this period. Previous work on the importance of subject choice during secondary school has focussed on whether young people study Science, Technology, Engineering and Maths (STEM)-related subjects (Tripney et al., 2010; Codiroli, 2015), particularly reflecting concerns about a gender gap in uptake of such subjects, although Codiroli (2015) highlights that this may not be the case among individuals from advantaged backgrounds. Particularly for science subjects, it seems plausible that universities are likely to prefer individuals who have taken these more detailed tracks. Just under a third (30%) of the sample report studying for at least two separate sciences or a double award.

When analysing whether individuals study foreign languages we only consider the main languages studied in English secondary schools: French, German, Italian and Spanish. In the data, all other subjects are simply encoded as 'Other' and we wish to exclude those who study for a qualification in their first language, which often makes up a majority of those studying such qualifications (Vidal Rodeiro, 2009). This cohort was one of the first for whom studying a language to age 16 was no longer compulsory; nevertheless, 60% study one of these main languages during this period. Finally, from the components of the EBacc, we compare individuals who study History or Geography with those who do not. Almost two thirds (64%) of the sample do so and, in common with other elements of the EBacc, they are generally more advantaged and have higher prior attainment than their peers who do not do so.

Table 3: Main reason young people report for choosing optional subjects at age 14

Reason	Percentage	Average Subject Selectivity Score
Advised by parent or teacher	4.0	-0.27
Subjects help or needed for next step	37.9	-0.05
Perceived ease of subjects for me	14.6	0.07
Enjoyment of subjects	36.8	0.10
Other reasons	6.8	-0.25
Total	100	0.00

Notes: Weighted using LSYPE Wave 2 sample design and non-response weights.

Next Steps also includes data on young people's self-reported reasons for choosing the optional subjects they pursue from age 14. We explore these reported motivations, while bearing in mind that self-reports should not necessarily be taken at face value. We group these as follows: advised by a parent of teachers; subjects help or are needed for their future plans; enjoyment of subjects; and unspecified other reasons. We report the frequency of these in Table 3, along with the average subject selective score of young people giving each response. 38% of young people report they primarily base this decision on the likely value of these subjects for they next steps in their education or entry to the labour market, while just slightly fewer (37%) report their main motivation is enjoyment of these subjects. Far fewer individuals report that they make their choice due to perceived ease of subjects for them (15%), because they were advised by their parent, teacher or school (4%) or other reasons (7%).

Individuals that report these different motivations for their choices also differ in the academic selectivity of the subjects they study. Perhaps surprisingly, those who say they choose subjects because they will help or are needed for their future plans are studying subjects with a below average academic selectivity. This appears to be because a larger proportion of individuals reporting this motivation are studying applied subjects, which they presumably see as helpful for non-academic tracks into the labour market. By contrast, those who report choosing their subjects on the basis of enjoyment choose the most academically selective mix of subjects.

Wave 7 of Next Steps covers young people aged 19-20. Hence the data allow us to model the entry to university through what might be thought of as the 'traditional' route, going from sixth form or further education college to university, either the same year or after a single gap year. While this includes the majority of those who attend university, later entrants would not be represented. The exclusion of this potentially interesting subpopulation should be noted; in particular, it could affect the results if subjects studied at GCSE are associated with later entry to university. We also consider entry to a Russell Group institution; the Russell Group is a group of 20 research-intensive UK institutions,[1]

[1]The Russell Group has since increased in size but for the individuals in the cohort considered it consisted of the following 20 universities: University of Birmingham, University of Bristol, University of Cambridge, Cardiff University, University of Edinburgh, University of Glasgow, Imperial College London, Kings College London, University of Leeds, University of

which are often considered to be amongst the most prestigious universities in the UK.

Next Steps includes a rich set of data measuring young people's socioeconomic status (SES), including household income, parental education, and parental occupational status, all of which are important in measuring SES (Hauser, 1994). Household income is measured at each wave between 1 and 4. As previous research has suggested 'permanent' income (rather than transitory income) has a much larger effect on young people's educational outcomes (Jenkins and Schluter, 2002, p.2). An approximation of the household's equivalised 'permanent' income is made by averaging across these four measures and dividing by the square root of household size. Previous work suggests that Next Steps underestimates household income to some extent, relative to social surveys where it is a major focus (Anders, 2012b).

Parental education also captures an important aspect of socioeconomic status; one explanation for this is that it "may alter the 'productivity' of [parents'] time investments in children" (Ermisch and Pronzato, 2010, p.1). Whatever the explanation, a number of studies have found evidence of a causal impact of parents education on children's educational outcomes (Chevalier, 2004; Ermisch and Pronzato, 2010; Havari and Savegnago, 2014), making it an important factor to take into account. Similarly, social class is seen by sociologists as a key element of an individual's SES (Goldthorpe and McKnight, 2004), in particular as "young people (and their families) have, as their major educational goal, the acquisition of a level of education that will allow them to attain a class position at least as good as that of their family of origin" (Breen and Yaish, 2006, p.232). Parents' occupational status is recorded in Next Steps using the National Statistics SocioEconomic Classification (NS-SEC), which was designed to capture social class differences between the different occupational types (Rose and Pevalin, 2001).

Table 4: Median family characteristics by quintile group of socioeconomic status index

Quintile group	Q1	Q2	Q3	Q4	Q5
Parental Education	< A*-C GCSE	A*-C GCSE	A Level	HE < Degree	Degree
Occupational Status	Routine occupations	Routine occupations	Intermediate occupations	Higher occupations	Higher occupations
Family Income (£p.a.)	5,830	9,780	13,286	16,618	29,910
N	1,680	1,525	1,580	1,618	1,602

Notes: Adjusted using LSYPE-provided Wave 7 survey design, attrition and non-response weights. Family income is equivalised by dividing by the square root of household size. Sample: Wave 7 respondents with non-missing data on university attendance, constituent socioeconomic indicators, subject choice variables, and prior attainment data.

The above measures are combined using principal components analysis with a polychoric correlation

Liverpool, London School of Economics and Political Science, University of Manchester, Newcastle University, University of Nottingham, University of Oxford, Queens University Belfast, University of Sheffield, University of Southampton, University College London and University of Warwick matrix (Olsson, 1979; Kolenikov and Angeles, 2009) to construct a single index of SES. Alternative methods, such as factor analysis, yield very similar results. This explains roughly three quarters of the variation in the three individual measures, but provides a broader measure of family circumstances than any one measure would provide. Table 4 reports the characteristics of the median member of quintile groups of this SES index.

4 Individual predictors of subject choices at age 14

This section aims to identify the extent to which socio-economic background, gender, ethnic group and school attended shape the patterns of subject and qualification choices at age 14. It makes a novel contribution in two important ways. First, it examines curriculum choices within compulsory education, before young people are able to select out of education so the sample of young people is heterogeneous and therefore represents a wider student body than studies that look at A Levels and university participation. Second, previous literature which has focused on participation in the individual GCSEs rather than the combination of all subjects chosen in combination.

We believe examining the patterns of subject choice in this way is more informative and the formation of our categories offers a unique metric to assess the subject choices made by 14-16 year olds. We compare a number of alternative approaches to describing the curriculum studied during this age range. We use the selectivity ranking of GCSE subjects defined according to average prior attainment discussed above as our primary outcome measure. In addition, we explore how participation in facilitating subjects, English Baccalaureate (EBacc)-eligible subjects, Science Technology Engineering and Maths (STEM) subjects and Applied GCSEs varies. This enables us to identify whether the social patterning of the curriculum varies according to alternative conceptions of a prestigious or highly-valued curriculum.

4.1 Methods

We employ a regression modelling approach, specifically making use of Ordinary Least Squares (OLS) regression models for the continuous selectivity measure and logistic regression models for our binary outcomes. We estimate sequential models which shed light on our three research questions in turn. Our first model only includes measures of individual and family background characteristics, providing insight on the conditional importance of these variables for individuals subject choices. The second model includes the same variables but also adds academic performance at age 14, thus shedding light on whether the inequalities we find descriptively and in the first model persist once prior attainment has been taken into account. Finally, our third model includes school-level characteristics, highlighting the role that such factors play.

We acknowledge that our modelling strategy is vulnerable to omitted variable bias, since our independent variables of interest, such as parental socio-economic status, are likely to be correlated with many individual- and school-level factors affecting a students ability, although we do try to minimise this issue through use of the rich background data (including prior attainment measures) available in Next Steps. As such, we do not view our results as truly causal, but rather capturing conditional relationships between background and subject choices. In addition, we account for the fact that observations are not truly independent from others attending the same school by calculating cluster-robust standard errors at school-level to conduct appropriate statistical inference.

4.2 Results

We begin by addressing the first research question regarding patterns of GCSE subjects taken according to social class, gender, ethnicity, parental education and income.

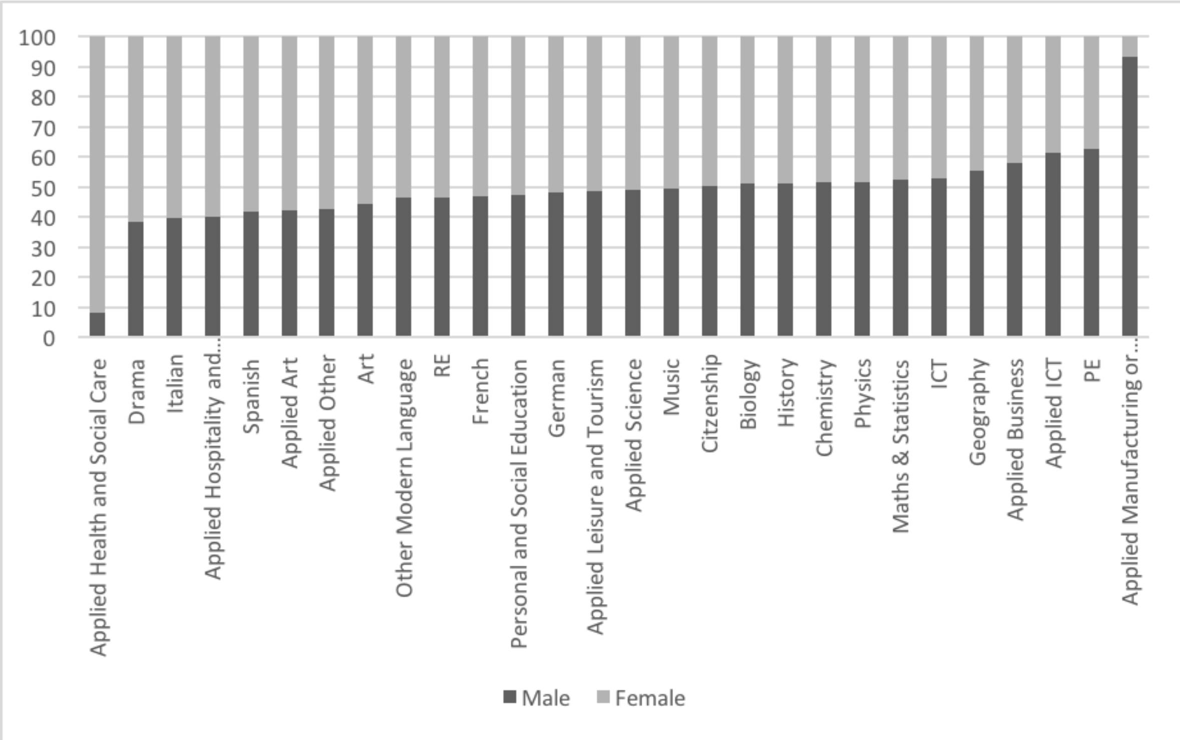
Figure 2 shows the gender differences in subjects chosen at GCSE. The widest gender gaps occur in Applied Manufacturing and Engineering which shows a higher uptake for boys and Applied Health and Social Care which shows a higher uptake for girls. Figure 3 shows the relationship between subject choice and class background. A higher proportion of young people from higher social class backgrounds take modern foreign languages including German, Italian, French and Spanish. In addition, we see larger proportion of young people with higher social class backgrounds taking music and the natural sciences, while young people from routine social class backgrounds are more often found in the applied subjects. A greater proportion of young people who are in the highest income quartiles do modern foreign languages, music and maths and statistics and a smaller proportion do applied subjects (Figure 4). As we might expect, the pattern is very similar for parental education as for income and social class (Figure 5). The story for subject choice by ethnicity shown in Figure 6 is not so clear. For example, there is no particular pattern evident in the subjects favoured by white students.

However, to address our second and third research question, we need to explore whether these patterns persist once we take into account prior attainment and school characteristics. To shed light on this issue we turn to the results of our regression modelling.

Table 2 reports coefficients from the OLS regression predicting the selectivity of the GCSE curriculum studied. As noted above, the capped selectivity score has been standardised to have a standard deviation of one. As such, coefficients may be interpreted as the expected change in standard deviations of the selectivity score associated with a one unit change in the independent variable.

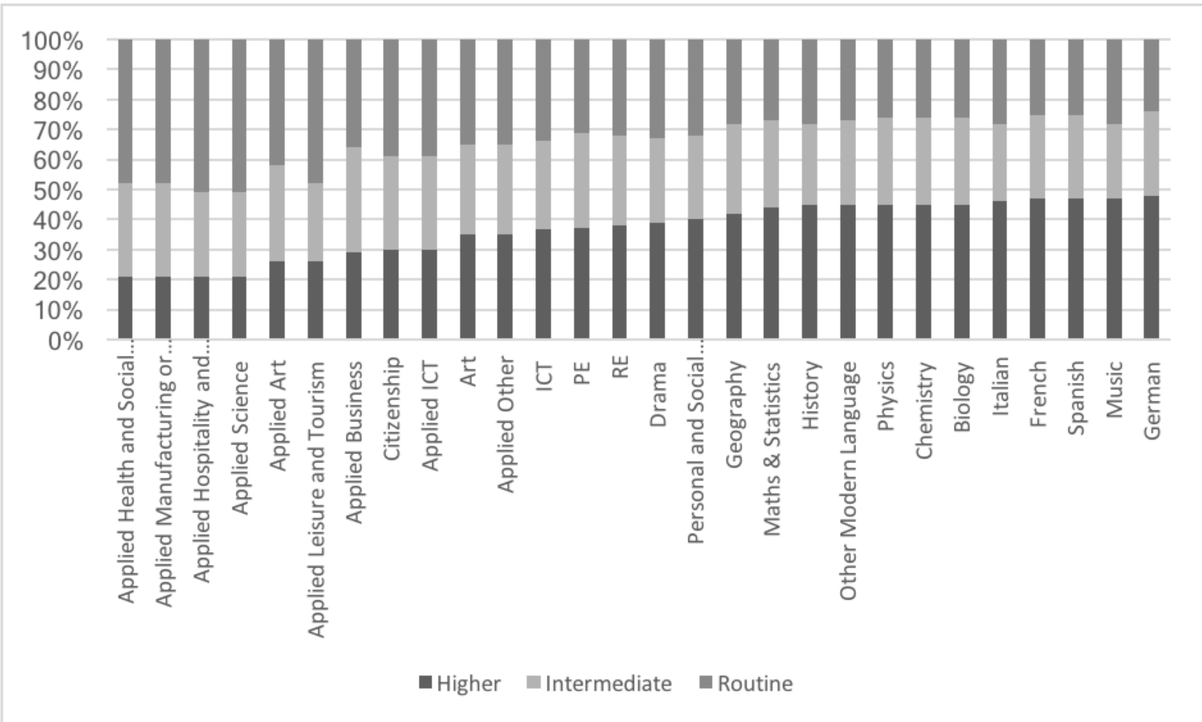
Model 1 reports changes in selectivity score associated with individual and familial characteristics, without controls for prior attainment or school attended. We find that students from wealthier backgrounds, and young people whose parents with higher levels of education study a more selective curriculum. Young people who live in rented accommodation also study less selective subjects then those who live in owner occupied or mortgaged property. There is also evidence that those from routine and intermediate backgrounds take a less selective curriculum than those from higher social class backgrounds. Girls study a more selective curriculum than boys, and young people with special educational needs take less-selective subjects. The only ethnicity coefficients which yield significant

Figure 2: Gender proportions of pupils studying each GCSE subject



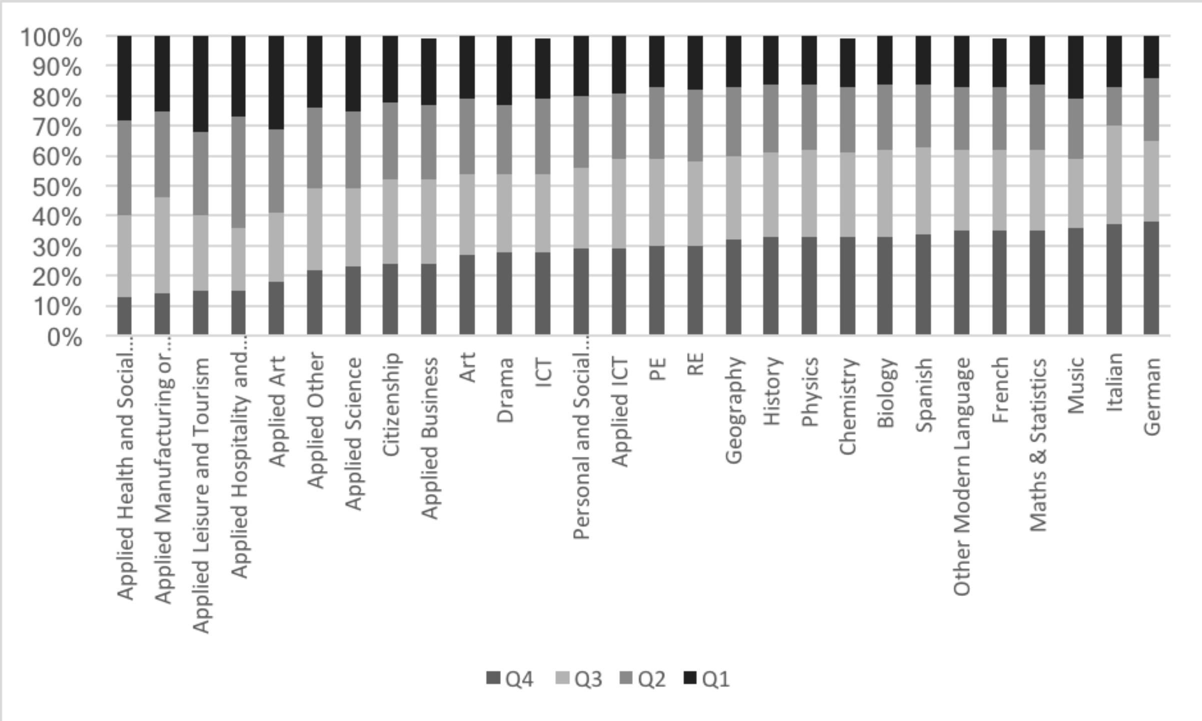
Notes: Adjusted using LSYPE-provided survey design and attrition weights. Sample: Wave 2 respondents with non-missing data on subject choice variables.

Figure 3: Parental occupational status proportions of pupils studying each GCSE subject



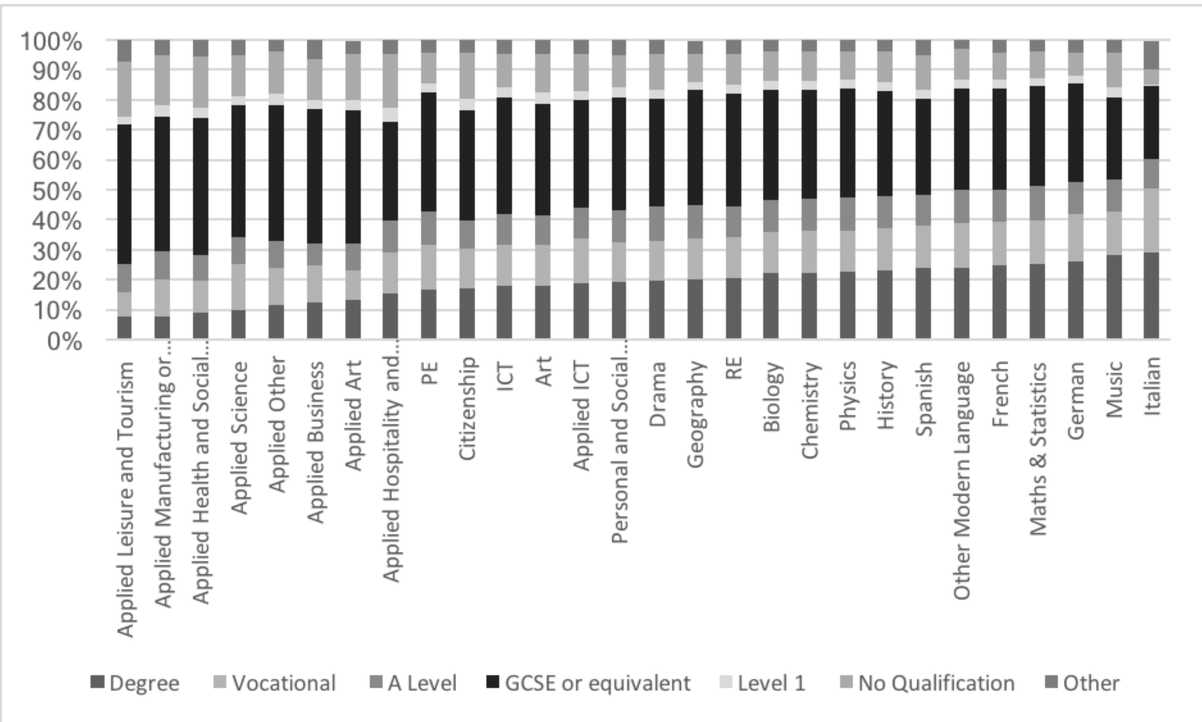
Notes: Adjusted using LSYPE-provided survey design and attrition weights. Sample: Wave 2 respondents with non-missing data on subject choice variables.

Figure 4: Parental income quintile group proportions of pupils studying each GCSE subject



Notes: Adjusted using LSYPE-provided survey design and attrition weights. Sample: Wave 2 respondents with non-missing data on subject choice variables.

Figure 5: Parental education group proportions of pupils studying each GCSE subject



Notes: Adjusted using LSYPE-provided survey design and attrition weights. Sample: Wave 2 respondents with non-missing data on subject choice variables.

Table 5: Estimated change in academic selectivity of subjects studied - linear regression model

	Model 1		Model 2		Model 3	
	β	SE	β	SE	β	SE
Ref. Higher Managerial						
Intermediate	-0.06*	(0.03)	-0.03	(0.03)	-0.04	(0.03)
Routine	-0.13***	(0.03)	-0.06*	(0.03)	-0.05*	(0.03)
Ref. Degree or Higher						
Other HE qualification	-0.17***	(0.03)	-0.08*	(0.03)	-0.07*	(0.03)
A Level	-0.17***	(0.04)	-0.08*	(0.03)	-0.06+	(0.03)
GCSE A-C	-0.32***	(0.03)	-0.16***	(0.03)	-0.14***	(0.03)
GCSE D-G and below	-0.33***	(0.04)	-0.12***	(0.03)	-0.09**	(0.03)
Income (per £10,000)	0.09***	(0.01)	0.04***	(0.01)	0.03**	(0.01)
Ref. Owns Property Outright/Mortgage						
Rent/Other	-0.13***	(0.02)	-0.05*	(0.02)	-0.02	(0.02)
Ref. White						
Mixed	-0.03	(0.04)	-0.03	(0.04)	-0.01	(0.04)
Indian	-0.04	(0.06)	-0.08	(0.05)	-0.05	(0.05)
Pakistani	-0.08	(0.06)	0.02	(0.06)	0.08	(0.06)
Bangladeshi	-0.15*	(0.07)	-0.15*	(0.07)	0.00	(0.08)
Black Caribbean	-0.07	(0.06)	0.04	(0.06)	0.09	(0.06)
Black African	-0.01	(0.07)	0.09	(0.06)	0.15*	(0.06)
Other	0.14**	(0.05)	0.12*	(0.05)	0.14**	(0.05)
Ref. Male						
Female	0.09***	(0.02)	0.07***	(0.02)	0.07***	(0.02)
Special Education Needs	-0.52***	(0.05)	-0.13*	(0.05)	-0.21***	(0.04)
Key Stage 3			0.05***	(0.00)	0.05***	(0.00)
Ref. State school						
Grammar School					0.27***	(0.07)
% FSM in school (SD)					-0.10***	(0.02)
Average Class Size					-0.08***	(0.02)
Ref. Co-ed						
Single-sex school					0.12*	(0.05)
Constant	0.26***	(0.05)	-1.65***	(0.09)	-1.52***	(0.08)
Observations	11,714		11,714		11,714	
Log likelihood	-15953		-15418		-15286	
DF	17		18		22	
Pseudo R-squared	0.09		0.17		0.19	

Notes: Sample: Wave 2 participants with valid data on subject choices and other model characteristics. Reporting regression coefficients. t statistics in parentheses. Stars indicate statistical significance: + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

results are the Bangladeshi and other ethnic groups, Bangladeshi pupils take less-selective subjects and other pupils take more selective subjects.

In Model 2, prior attainment at age 14 is added to the model. Unsurprisingly, prior attainment is positively and significantly associated with the selectivity of subjects selected. Controlling for prior attainment, the differentials due to socio-economic factors are reduced, but significant differences remain. The ethnic and gender differences are not accounted for by prior attainment.

Once school characteristics are included in Model 3, the results do not change substantively. The class, parental education, income, gender and prior attainment coefficients remain statistically significant. However, housing tenure is no longer significant and the ethnicity coefficients change. We see that black African young people take more-selective subjects, once school characteristics are controlled for. Some of the school variables are also of interest, in particular we find that young people who attend grammar schools take a more selective curriculum, so too do those who attend a single-sex schools. Larger class size is associated with a decrease in the selectivity of the curricula, as is a higher proportion of free-school meal eligible students in the school.

Table 3 shows the results predicting taking three or more facilitating subjects. Odds ratios are presented to identify the relative importance of exposure to various predictors. An odds ratio of one means that the exposure to a given predictor does not affect the outcome; an odds ratio of greater than one means higher odds of the outcome, while an odds ratio of less than one means lower odds of an outcome. Model 1 shows that equivalised household income is positively associated with taking three or more facilitating subjects, there is also evidence from this model that young people from lower social class backgrounds have lower odds, so too do young people whose parents have lower levels of education. Living in rented accommodation is associated with lower odds of taking facilitating subjects. With respect to ethnicity, there is some variation, namely that mixed race young people and black Caribbean young people have lower odds, while Indian and other young people have higher odds of taking facilitating subjects.

Model 2 shows that prior attainment is positively associated with taking three or more facilitating subjects and accounting for this renders the social class difference non-significant, but significant differences remain according to parental education, income and ethnic group. The inclusion of prior attainment renders the coefficient for black Caribbean insignificant and the Pakistani and Black African groups positively associated and significant. When school characteristics are included in Model 3, some of the results change further, for example Bangladeshi young people are significantly more likely than whites to take three or more facilitating subjects. Income and prior attainment remain significant predictors for taking three or more facilitating subjects, controlling for school characteristics. Attending a grammar school is weakly associated with an increase in odds of doing three or

Table 6: Estimated change in the odds of studying at least three facilitating subjects - logistic regression model

	Model 1		Model 2		Model 3	
	OR	SE	OR	SE	OR	SE
Ref. Higher Managerial						
Intermediate	0.98	(0.06)	1.05	(0.07)	1.04	(0.07)
Routine	0.83**	(0.06)	0.98	(0.07)	1.00	(0.07)
Ref. Degree or Higher						
Other HE qualification	0.69***	(0.06)	0.84*	(0.07)	0.84*	(0.07)
A Level	0.69***	(0.06)	0.85+	(0.07)	0.84+	(0.07)
GCSE A-C	0.51***	(0.04)	0.73***	(0.05)	0.74***	(0.06)
GCSE D-G and below	0.48***	(0.04)	0.77**	(0.06)	0.80*	(0.07)
Income (per £10,000)	1.20***	(0.04)	1.12***	(0.03)	1.08**	(0.03)
Ref. Owns Property Outright/Mortgage						
Rent/Other	0.77***	(0.04)	0.92+	(0.05)	0.95	(0.05)
Ref. White						
Mixed	0.80*	(0.08)	0.79*	(0.08)	0.83+	(0.09)
Indian	1.46**	(0.18)	1.38*	(0.18)	1.48**	(0.20)
Pakistani	1.18	(0.14)	1.55***	(0.18)	1.83***	(0.23)
Bangladeshi	1.17	(0.15)	1.18	(0.15)	1.57**	(0.23)
Black Caribbean	0.73*	(0.09)	0.93	(0.12)	1.04	(0.14)
Black African	1.13	(0.15)	1.47**	(0.20)	1.68***	(0.24)
Other	1.31*	(0.15)	1.27*	(0.15)	1.39**	(0.17)
Ref. Male						
Female	0.99	(0.04)	0.93	(0.04)	0.93+	(0.04)
Special Education Needs	0.35***	(0.03)	0.87	(0.08)	0.83*	(0.08)
Key Stage 3			1.14***	(0.01)	1.13***	(0.01)
Ref. State school						
Grammar School					1.40+	(0.27)
% FSM in school (SD)					0.82***	(0.03)
Average Class Size					0.99	(0.04)
Ref. Co-ed						
Single-sex school					1.18	(0.12)
Constant	2.49***	(0.28)	0.02***	(0.00)	0.03***	(0.01)
Observations	11,714		11,714		11,714	
Log likelihood	-7733		-6761		-6728	
DF	17		18		22	
Pseudo R-squared	0.06		0.13		0.14	

Notes: Sample: Wave 2 participants with valid data on subject choices and other model characteristics. Reporting odds ratios. t statistics in parentheses. Stars indicate statistical significance: + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

more facilitating subjects, while single-sex schooling is not significantly different from co-educational schooling. As the proportion of FSM eligible students in the school increases, the odds of doing three or more facilitating subjects decreases. There is no significant effect found for average class sizes.

Table 4 shows our models of whether the young person takes the EBacc. In Model 1, parental education, income and housing tenure are all significant predictors of taking EBacc-eligble subjects. Black Caribbean and mixed race young people have lower odds of taking the EBacc while other have higher odds compared to white pupils. There is no significant gender difference.

Once prior attainment at key stage 3 is controlled in Model 2, parental education, income and housing tenure remain significant. The pattern according to ethnic group changes in this model, as a positive differential in favour of Pakistani pupils emerges compared to whites of the same level of prior attainment. In this model, once prior attainment is accounted for, the odds of girls doing EBacc-eligble subjects is significantly lower than for boys.

Once school characteristics are controlled in Model 3, the income, parental education and prior ability coefficients remain significant. The Pakistani advantage is increased. Attending a grammar school or a single-sex school increases the odds of taking the EBacc. As the proportion of FSM-eligible students in the school increases, the odds of doing EBacc-eligible subjects declines.

Table 5 shows our models predicting the odds of taking three or more STEM subjects at GCSE. Model 1 shows that household income, home ownership and higher parental education increase the odds of taking three STEM subjects. Black Caribbean, black African and mixed race pupils all have reduced odds of taking three STEM subjects. As we expect, girls have lower odds of doing three or more STEM subjects than boys.

Model 2 shows that the inclusion of prior attainment predicts selection of three or more STEM subjects positively. Prior attainment fully explains the parental education and housing tenure differentials, and the income difference is also no longer significant. In other words, socio-economic differentials in access to STEM are largely driven by prior attainment. With the exception of mixed race young people, the ethnic differences become non-significant, suggesting that the apparent disadvantage experienced by these groups regarding STEM can also be accounted for by prior attainment. In contrast, the gender differences remain.

The negative association for girls remains once school characteristics are included in Model 3, so too does the negative association for mixed race young people compared to white young people. In contrast, the odds for Bangladeshi young people become significantly positive once school characteristics are controlled. Participation in STEM subjects does not vary by school characteristics, with

Table 7: Estimated change in the odds of studying EBacc subjects - logistic regression model

	Model 1		Model 2		Model 3	
	OR	SE	OR	SE	OR	SE
Ref. Higher Managerial						
Intermediate	0.90+	(0.05)	0.96	(0.06)	0.94	(0.06)
Routine	0.77***	(0.05)	0.89+	(0.06)	0.91	(0.06)
Ref. Degree or Higher						
Other HE qualification	0.71***	(0.05)	0.87+	(0.07)	0.87+	(0.07)
A Level	0.68***	(0.06)	0.83*	(0.07)	0.83*	(0.07)
GCSE A-C	0.53***	(0.04)	0.76***	(0.05)	0.78***	(0.05)
GCSE D-G and below	0.51***	(0.04)	0.79**	(0.06)	0.84*	(0.07)
Income (per £10,000)	1.22***	(0.04)	1.10***	(0.03)	1.07**	(0.03)
Ref. Owns Property Outright/Mortgage						
Rent/Other	0.75***	(0.04)	0.89*	(0.05)	0.93	(0.05)
Ref. White						
Mixed	0.81*	(0.08)	0.81*	(0.09)	0.84	(0.09)
Indian	1.09	(0.13)	1.01	(0.13)	1.08	(0.14)
Pakistani	1.07	(0.12)	1.36*	(0.16)	1.62***	(0.21)
Bangladeshi	0.87	(0.12)	0.85	(0.12)	1.19	(0.19)
Black Caribbean	0.62**	(0.09)	0.79	(0.12)	0.88	(0.14)
Black African	0.87	(0.12)	1.09	(0.15)	1.24	(0.17)
Other	1.24*	(0.14)	1.19	(0.13)	1.30*	(0.15)
Ref. Male						
Female	0.94	(0.05)	0.90*	(0.04)	0.91+	(0.04)
Special Education Needs	0.43***	(0.04)	0.99	(0.11)	0.92	(0.10)
Key Stage 3			1.13***	(0.01)	1.12***	(0.01)
Ref. State school						
Grammar School					1.95**	(0.43)
% FSM in school (SD)					0.78***	(0.04)
Average Class Size					0.95	(0.04)
Ref. Co-ed						
Single-sex school					1.26*	(0.15)
Constant	0.89	(0.09)	0.01***	(0.00)	0.01***	(0.00)
Observations	11,714		11,714		11,714	
Log likelihood	-7147		-6671		-6597	
DF	17		18		22	
Pseudo R-squared	0.06		0.12		0.13	

Notes: Sample: Wave 2 participants with valid data on subject choices and other model characteristics. Reporting odds ratios. t statistics in parentheses. Stars indicate statistical significance: + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 8: Estimated change in the odds of studying STEM subjects - logistic regression model

	Model 1		Model 2		Model 3	
	OR	SE	OR	SE	OR	SE
Ref. Higher Managerial						
Intermediate	0.96	(0.06)	0.99	(0.06)	0.98	(0.06)
Routine	0.82**	(0.05)	0.89+	(0.06)	0.91	(0.06)
Ref. Degree or Higher						
Other HE qualification	0.95	(0.07)	1.06	(0.08)	1.07	(0.08)
A Level	0.82*	(0.07)	0.91	(0.08)	0.92	(0.08)
GCSE A-C	0.77***	(0.06)	0.94	(0.07)	0.96	(0.07)
GCSE D-G and below	0.74***	(0.06)	0.95	(0.08)	1.00	(0.08)
Income (per £10,000)	1.08**	(0.03)	1.03	(0.03)	1.00	(0.03)
Ref. Owns Property Outright/Mortgage						
Rent/Other	0.89*	(0.05)	0.97	(0.05)	1.02	(0.05)
Ref. White						
Mixed	0.74**	(0.08)	0.74**	(0.08)	0.77**	(0.08)
Indian	1.06	(0.13)	1.01	(0.13)	1.08	(0.14)
Pakistani	0.85	(0.11)	0.95	(0.12)	1.11	(0.14)
Bangladeshi	1.06	(0.17)	1.06	(0.17)	1.41*	(0.22)
Black Caribbean	0.71*	(0.10)	0.81	(0.11)	0.90	(0.12)
Black African	0.75*	(0.11)	0.83	(0.12)	0.93	(0.13)
Other	1.02	(0.13)	0.99	(0.13)	1.08	(0.14)
Ref. Male						
Female	0.72***	(0.04)	0.70***	(0.04)	0.70***	(0.04)
Special Education Needs	0.56***	(0.06)	0.87	(0.09)	0.85+	(0.09)
Key Stage 3			1.06***	(0.01)	1.05***	(0.00)
Ref. State school						
Grammar School					1.43	(0.40)
% FSM in school (SD)					0.81***	(0.04)
Average Class Size					1.02	(0.05)
Ref. Co-ed						
Single-sex school					1.15	(0.16)
Constant	0.75**	(0.08)	0.08***	(0.02)	0.10***	(0.02)
Observations	11,714		11,714		11,714	
Log likelihood	-7200		-7070		-7028	
DF	17		18		22	
Pseudo R-squared	0.02		0.04		0.05	

Notes: Sample: Wave 2 participants with valid data on subject choices and other model characteristics. Reporting odds ratios. t statistics in parentheses. Stars indicate statistical significance: + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

the exception of the proportion of FSM in the school which is negatively associated with doing three or more STEM subjects.

Table 6 shows the results predicting taking one or more applied GCSEs. In Model 1 we observe that social class is a significant predictor for studying applied GCSEs, and those from routine backgrounds are more likely to study these subjects. We also observe that children in more affluent households have lower odds of studying any applied subjects. Parental education is also significant: as the highest level of parental education decreases the odds of the young person studying applied GCSEs increase. Furthermore, we see that girls are more likely than boys to study applied GCSEs, so too are those with special education needs.

Model 2 includes prior educational attainment and the significant odds ratio of less than one suggests that as ability increases, the odds of doing applied GCSEs reduces. The inclusion of prior attainment explains the social class differential in taking applied subjects. None of the other significant variables from Model 1 change once prior attainment is taken into account, except that the apparent influence of parental education reduces slightly and special educational needs become associated with lower odds of taking one or more applied GCSEs.

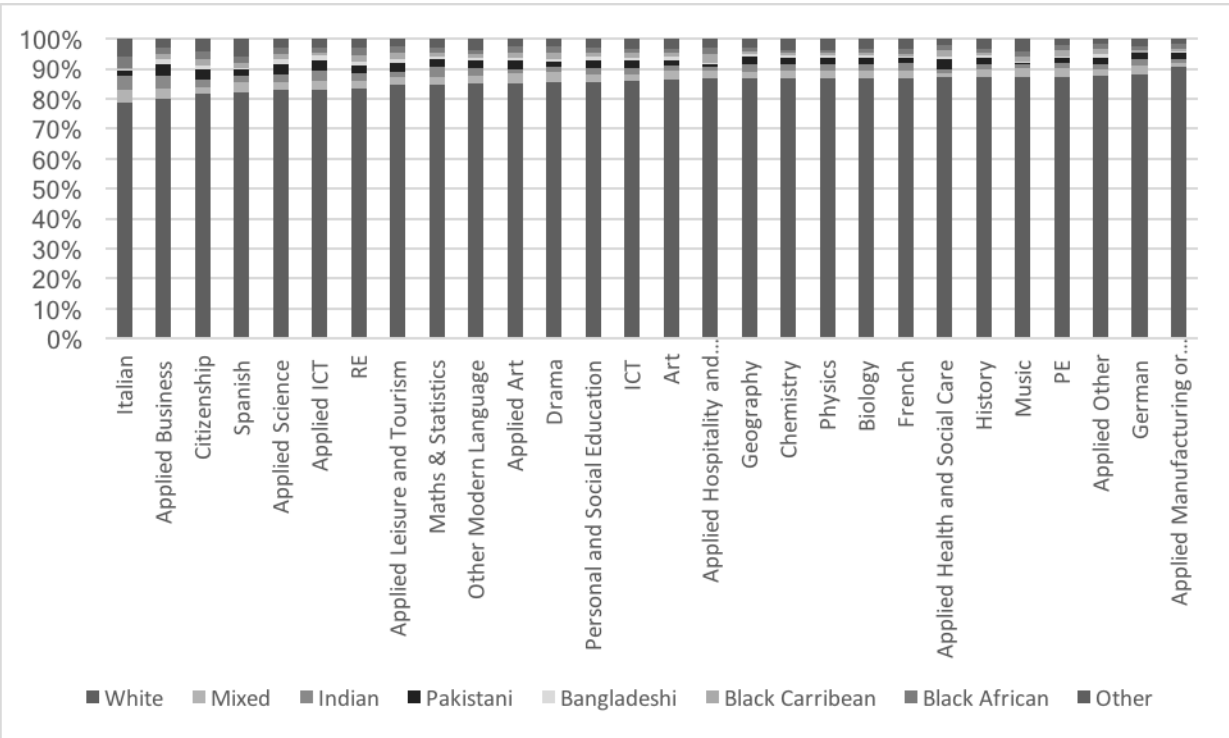
Once school characteristics are included in Model 3 the variables do not change substantively. We observe that young people who attend grammar schools have significantly lower odds of taking applied GCSEs compared with those in non-selective schools. Furthermore, attending a single-sex school is associated with lower odds of studying applied subjects over and above individual characteristics.

Table 9: Estimated change in the odds of studying any applied subjects - logistic regression model

	Model 1		Model 2		Model 3	
	OR	SE	OR	SE	OR	SE
Ref. Higher Managerial						
Intermediate	1.11	(0.07)	1.07	(0.07)	1.08	(0.07)
Routine	1.16*	(0.07)	1.06	(0.07)	1.07	(0.07)
Ref. Degree or Higher						
Other HE qualification	1.33***	(0.10)	1.19*	(0.09)	1.17*	(0.09)
A Level	1.30**	(0.11)	1.16+	(0.09)	1.14	(0.09)
GCSE A-C	1.73***	(0.12)	1.42***	(0.10)	1.39***	(0.10)
GCSE D-G and below	1.56***	(0.12)	1.21*	(0.09)	1.22*	(0.10)
Income (per £10,000)	0.84***	(0.02)	0.89***	(0.03)	0.90***	(0.03)
Ref. Owns Property Outright/Mortgage						
Rent/Other	1.13*	(0.06)	1.02	(0.05)	1.04	(0.05)
Ref. White						
Mixed	0.89	(0.09)	0.89	(0.09)	0.92	(0.09)
Indian	0.99	(0.10)	1.04	(0.10)	1.09	(0.11)
Pakistani	1.11	(0.12)	0.99	(0.11)	1.10	(0.12)
Bangladeshi	1.00	(0.15)	1.00	(0.16)	1.16	(0.17)
Black Caribbean	1.10	(0.14)	0.96	(0.12)	1.06	(0.14)
Black African	0.95	(0.12)	0.84	(0.11)	0.92	(0.12)
Other	0.70**	(0.09)	0.71**	(0.09)	0.78*	(0.10)
Ref. Male						
Female	1.11*	(0.05)	1.14**	(0.05)	1.14**	(0.05)
Special Education Needs	1.25*	(0.11)	0.76**	(0.08)	0.82*	(0.07)
Key Stage 3			0.94***	(0.00)	0.94***	(0.00)
Ref. State school						
Grammar School					0.37***	(0.11)
% FSM in school (SD)					0.94	(0.04)
Average Class Size					1.07+	(0.04)
Ref. Co-ed						
Single-sex school					0.76*	(0.09)
Constant	0.64***	(0.07)	7.24***	(1.33)	6.15***	(1.08)
Observations	11,714		11,714		11,714	
Log likelihood	-7888		-7713		-7654	
DF	17		18		22	
Pseudo R-squared	0.03		0.05		0.06	

Notes: Sample: Wave 2 participants with valid data on subject choices and other model characteristics. Reporting odds ratios. t statistics in parentheses. Stars indicate statistical significance: + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Figure 6: Ethnic group proportions of pupils studying each GCSE subject



Notes: Adjusted using LSYPE-provided survey design and attrition weights. Sample: Wave 2 respondents with non-missing data on subject choice variables.

5 The role of schools in individuals subject choices at age 14

This section explores the extent to which the subjects that individuals study from age 14-16 are associated, not only with their personal characteristics, but also with the school they attend. It also builds on this, to provide evidence about the extent to which schools' provision (and, hence, the options open to their pupils) is driven by their composition in terms of prior academic attainment, socioeconomic background and gender mix. These are all factors that previous work has found to influence the subjects that individuals study (Davies et al., 2008), but given Jin et al.'s (2011) observation that the most common reason an individual reports they cannot study a subject is that it is not available, it seems important to explore whether it is indeed due to what schools can feasibly offer.

After an initial descriptive exploration, we apply hierarchical regression modelling in order to explore how variance is partitioned between- and within-schools, conditional on these other factors. The main results are estimated using administrative data from the National Pupil Database. However, we also demonstrate the robustness of the results to use of richer measures of individuals' socioeconomic status by replicating our broad findings using survey data from Next Steps (the Longitudinal Study of Young People in England) as far as possible. Overall, our results demonstrate the important role that schools seem to play in subject choice decisions, with significant variation in subjects studied attributable to the school-level. However, they also highlight the constraints that schools face from what they can viably offer, with a significant proportion of this school-level variation being explained by school composition and differences in subjects studied among pupils in non-selective schools in local areas with selective schooling.

5.1 Hierarchical regression modelling

In order to explore the variation in subject choice in a more formal setting, attempting to decompose the influence of prior attainment, SES and gender, we fit hierarchical regression models of the outcomes of interest. Hierarchical models (also known as multi-level models or regression models with random effects) explicitly partition the variance in the model between units and within units (in this case schools). Marks (2006) uses a similar technique when looking at within- and between-school variation in academic performance.

Hierarchical regression models of the following form are estimated. For continuous outcome variable (index of subject selectivity):

$$y_{ij} = \beta_0 + \beta X'_{ij} + \gamma \bar{X}'_j + \eta_j + \varepsilon_{ij} \tag{1}$$

For dichotomous outcome variables:

$$ln(\frac{y_{ij}}{1-y_{ij}}) = \beta_0 + \beta X'_{ij} + \gamma \bar{X}'_j + \eta_j + \varepsilon_{ij} \quad (2)$$

where y is the outcome variable, i indicates the individual, j indicates the school which they attend, X represents a vector of individual-level regressors, $\bar{X}$ represents a vector of the school-level averages of the individual-level regressors, and β and γ represent fixed-effects coefficients. Finally, there are two error terms: η represents a normally distributed school-level random intercept, and ε represents an individual-level error term (normally distributed in the case of models with a continuous outcome and with a logistic distribution in the case of models with a dichotomous outcome).

Models including random effects make assumptions about relationships in the data, which are typically unlikely to be strictly justified. First, it is assumed that the school-level effects are normally distributed. Second, it is assumed that there is no correlation between the individual-level and the school-level error terms. However, one strategy for relaxing this latter assumption is modelling this relationship by including group-level means of the regressors (Mundlak, 1978). Given their substantive interest, these are being included in the model in any case. As such, this gives us increased confidence in the estimated relationships between the individual-level covariates and the outcome variables.

Given our interest in the role of schools, an important element of the results is the proportion of variance explained simply by the school which an individual attends i.e. not the proportion explained by within-school variation or the proportion explained by school-level variation in the make-up of the pupils. This is known as the intra-cluster correlation (ICC or ρ) and is calculated from the hierarchical models as follows:

$$\rho = \frac{var(\eta)}{var(\eta) + var(\varepsilon)} \quad (3)$$

Covariates included in the model are added in a sequential manner as follows. First, an empty model (M0) is estimated. This performs the important function of providing a baseline unconditional intra-cluster correlation against which the conditional intra-cluster correlations in later models may be compared.

In the first model including covariates (M1), we add an individual's KS2 standardised score (Z-Score) as well as the school's average KS2 Z-Score. This provides results on the conditional association between young people's prior attainment after controlling for the school's performance. It aims to capture not only whether individuals with higher prior attainment study more academically selective

Table 10: Distribution of pupils from each quintile group of KS2 performance and the quintile group of the school's intake as measured by KS2

School / Individual	Q1 (Low)	Q2	Q3	Q4	Q5 (High)	Total	Group Size
Q1 (Low)	0.36	0.24	0.19	0.13	0.08	1.00	(0.09)
Q2	0.26	0.23	0.21	0.18	0.12	1.00	(0.17)
Q3	0.20	0.22	0.21	0.20	0.17	1.00	(0.22)
Q4	0.15	0.19	0.22	0.22	0.22	1.00	(0.27)
Q5 (High)	0.07	0.12	0.17	0.25	0.40	1.00	(0.25)
Overall	0.18	0.19	0.20	0.21	0.22	1.00	(1.00)

Notes: Reporting row proportions, except for final column which reports column proportions. Sample: Young people at state-schools with valid data on subject choices, gender, SES, and KS2 performance. Sample size = 542,944.

subjects but also whether there is an additional increase in the average subject mix of individuals in schools with an intake of more highly attaining pupils.

The next two models (M2 and M3) adds in covariates related to gender and SES, respectively; as with M1 these capture both whether these individual-level characteristics are relevant for subjects studied and whether the school context in these terms has additional predictive power. In M2 we include a dummy variable for an individual's gender. We allow for a more flexible relationship between the school's gender balance and subject choice a simple linear relationship. We add dummy variables that categorise schools into gender-balanced (our baseline), mainly male, mainly female, all-male, and all-female. M3 adds the standardised index of young people's SES, along with the school's average value of this index. This provides evidence on whether there is an independent association between SES and subject choice once prior attainment has been held constant.

Our final model (M4) looks specifically at a possible constraint on schools. We include an indicator for whether individuals are in a school located in a Local Authority (LA) in which more than 5% of pupils attend selective (grammar) schools and for whether they are in a selective school themselves. While much of the dynamic of being in a school with high prior attainment will be captured through the covariates introduced in M1, we are particularly interested to see whether being in a non-selective school within an area in which selective schools are present affects the subjects studied.

5.2 Results

Given the focus of this paper on the variation between schools the set of independent variables to be used in the model is deliberately parsimonious.

We include a young people's academic attainment at age 11. This age is used since it is at this point that young people sort into secondary schools. The attainment measure at age 11 is based on young people's performance in Key Stage 2 tests in English, maths and science. We standardise this variable to be a Z-score (i.e. mean of zero, standard deviation of one) to aid interpretation. The

Table 11: Distribution of pupils from each quintile group of SES and the quintile group of the school's SES intake

School / Individual	Q1 (Low)	Q2	Q3	Q4	Q5 (High)	Total	Group Size
Q1 (Low)	0.61	0.24	0.09	0.04	0.02	1.00	(0.10)
Q2	0.28	0.32	0.21	0.12	0.07	1.00	(0.16)
Q3	0.13	0.25	0.26	0.21	0.15	1.00	(0.21)
Q4	0.06	0.15	0.25	0.29	0.25	1.00	(0.26)
Q5 (High)	0.02	0.07	0.16	0.29	0.46	1.00	(0.27)
Overall	0.16	0.18	0.21	0.22	0.23	1.00	(1.00)

Notes: Reporting row proportions, except for final column which reports column proportions. Sample: Young people at state-schools with valid data on subject choices, gender, SES, and KS2 performance. Sample size = 542,944.

Table 12: Distribution of pupils by gender and group of the school's gender distribution

School / Individual	Female	Male	Overall	Group Size
All Female	1.00		1.00	(0.08)
Mainly Female	0.58	0.42	1.00	(0.18)
Gender Balanced	0.51	0.49	1.00	(0.60)
Mainly Male	0.41	0.59	1.00	(0.08)
All Male		1.00	1.00	(0.05)
Overall	0.52	0.48	1.00	(1.00)

Notes: Reporting row proportions, except for final column which reports column proportions. Sample: Young people at state-schools with valid data on subject choices, gender, SES, and KS2 performance. Sample size = 542,944. School gender distribution groups are as follows: All female: Prop. male $= 0.00$; Mainly female: $0.00 <$ Prop. male 0.45; Gender balanced: $0.45 \leq$ Prop. male ≤ 0.55; Mainly male: $0.55 <$ Prop. male < 1.00; All male: Prop. male = 1.00. Group sizes may not quite sum to 1 due to rounding.

distribution of pupils by quintile group of their KS2 performance and the quintile group of their school's KS2 intake is reported in Table 10. While schools with low KS2 intakes have, by definition, larger numbers of pupils whose KS2 performance is in the bottom quintile group, there are still individuals in such schools with high levels of KS2 performance.

The NPD includes two proxies for individuals' socioeconomic status. These are combined using principal components analysis with a polychoric correlation matrix (Olsson, 1979; Kolenikov and Angeles, 2009) to construct a single index of SES (alternative methods, such as factor analysis, yield very similar results.). This explains roughly three quarters of the variation in the two individual measures. Again, we ensure that this has a mean of zero and a standard deviation of one. There is a similar pattern of the SES distribution within and between schools as that described above regarding prior attainment (Table 11).

On gender, we find that 60% of pupils are in broadly gender balanced schools, while 13% are in single gender schools (8% female-only and 5% male only), 18% are in mainly female schools, and 8% are in mainly male schools.

We begin with a descriptive exploration of differences in young people's subject choices depending not only on their personal characteristics, but also on the composition of their school in terms of prior

Table 13: Average academic subject selectivity of individuals by their KS2 performance and the average KS2 performance of their schoolmates

School / Individual	Q1 (Low)	Q2	Q3	Q4	Q5 (High)	Overall
Q1 (Low)	-0.77	-0.50	-0.33	-0.15	0.04	-0.59
Q2	-0.66	-0.36	-0.18	-0.01	0.20	-0.24
Q3	-0.54	-0.22	-0.03	0.15	0.37	-0.02
Q4	-0.50	-0.16	0.03	0.23	0.44	0.21
Q5 (High)	-0.38	-0.07	0.17	0.39	0.66	0.49
Overall	-0.48	-0.27	-0.07	0.05	0.35	0.00

Notes: Cells report average standardised capped academic subject selectivity. Sample: Young people at state-schools with valid data on subject choices, gender, SES, and KS2 performance. Sample size = 542,944.

Table 14: Average academic subject selectivity of individuals by their SES background and the average SES of their schoolmates

School / Individual	Q1 (Low)	Q2	Q3	Q4	Q5 (High)	Overall
Q1 (Low)	-0.51	-0.33	-0.20	-0.26	-0.14	-0.44
Q2	-0.49	-0.28	-0.18	-0.09	0.00	-0.17
Q3	-0.38	-0.19	-0.06	0.06	0.08	0.03
Q4	-0.25	0.00	0.12	0.19	0.29	0.15
Q5 (High)	-0.21	0.02	0.19	0.25	0.33	0.26
Overall	-0.42	-0.27	-0.09	0.14	0.25	0.00

Notes: Cells report average standardised capped academic subject selectivity. Sample: Young people at state-schools with valid data on subject choices, gender, SES, and KS2 performance. Sample size = 542,944.

attainment, socioeconomic status, and gender. In the interests of space, this descriptive exploration focuses only on our subject academic selectivity measure; we broaden our focus to other specific subject choices using regression models below.

Average academic subject selectivity is increasing in both an individual's prior attainment and the average prior attainment of their schoolmates (Table 13). However, this is not just because the school-level association picks the average change in individuals' prior attainment within the school; we can also see that those with low prior attainment but in high prior attainment intake schools are more likely to study more academically selective subjects than their counterparts in lower prior attainment schools.

There is a similar overall pattern when it comes to socioeconomic status, perhaps in part because there is a correlation ($r = 0.28$) between the measures of SES and performance at KS2. On an individual level, SES is, reassuringly, less predictive of the academic selectivity of subjects studied than young people's prior attainment. Notably, the SES intake of a school appears just as important a predictor as the individual-level indicator.

Academic selectivity of subjects studied also varies somewhat by gender, with females studying a set of courses with a slightly higher academic selectivity, on average (Table 15). However, in this case, the selectivity of subjects studied varies much more by the gender composition of their school.

Table 15: Average academic subject selectivity of individuals by their gender and the gender composition of schoolmates

School / Individual	Female	Male	Overall
All Female	0.31	-	0.31
Mainly Female	-0.06	-0.19	-0.11
Gender Balanced	0.02	-0.09	-0.03
Mainly Male	-0.00	-0.06	-0.04
All Male	-	0.36	0.36
Overall	0.05	-0.06	0.00

Notes: Cells report average standardised capped academic subject selectivity. Sample: Young people at state-schools with valid data on subject choices, gender, SES, and KS2 performance. Sample size = 542,944.

Most noticeably, pupils in single sex schools study a set of subjects of significantly higher academic selectivity than those in mixed schools.

However, it is important to recognise that the analysis above only attempts to explore one aspect of the association between young people's characteristics, schools' composition and the subjects studied by its pupils at a time. There are reasons to think that some of the associations are due to other confounding factors, especially in relatively small groups such as single-sex schools. In order to alleviate this issue to some extent, and to allow us to explore the importance of these factors for specific subject choices in a manageable way, we now turn to hierarchical regression modelling.

The results tables focus on the conditional associations between individual-level characteristics and subject choices, school-level compositions and subject choices, and the proportion of variance explained at the school level conditional on included covariates (ρ). For models of the continuous index of academic selectivity we report the regression coefficients, which are interpreted as the expected change in the subject selectivity index in standard deviations.

We consider the results for the subject selectivity score first (Table 16). In M0, in which there are no regressors, the only estimate to be interpreted is the proportion of variance between schools rather than within schools, which is estimated to be 0.32. This suggests that approximately two thirds of the variation in individuals' subject selectivity is between different pupils within the same school. After conditioning on school intake in terms of prior attainment (M1) this proportion of variation explained by schools is reduced to 0.27 in models M1; this is only reduced marginally further after conditioning on gender or socioeconomic status in M2 and M3.

Turning to associations between characteristics and subject choices, on adding controls for prior attainment at age 11 (M1), we find that a one standard deviation increase in an individual's prior attainment is associated with roughly a 0.3 of a standard deviation change in subject selectivity score. The school's intake attainment profile is associated with subject selectivity even more strongly, with a one standard deviation change in school intake being associated with 0.4 of a standard deviation

Table 16: Variation in subject selectivity

	M0	M1	M2	M3	M4
KS2 Z-Score		0.321	0.322	0.308	0.308
		(76.93)***	(77.21)***	(76.10)***	(76.10)***
School KS2 Z-Score		0.420	0.379	0.193	0.172
		(17.48)***	(14.23)***	(6.30)***	(4.14)***
Male			-0.127	-0.130	-0.130
			(-24.13)***	(-24.77)***	(-24.77)***
All-male school			0.156	0.270	0.278
			(3.91)***	(6.37)***	(6.35)***
Mainly-male school			0.0302	0.0998	0.105
			(0.83)	(2.77)**	(2.91)**
Mainly-female school			-0.0722	-0.0611	-0.0604
			(-2.60)**	(-2.27)*	(-2.24)*
All-female school			0.103	0.230	0.239
			(3.22)**	(6.70)***	(6.75)***
SES Z-Score				0.107	0.107
				(39.75)***	(39.75)***
School SES Z-Score				0.124	0.129
				(4.90)***	(5.01)***
$>$5% in LA in selective schools					-0.0603
					(-2.00)*
Selective school					0.0659
					(1.22)
Dep. Var. Mean	0.00	0.00	0.00	0.00	0.00
ρ	0.32	0.27	0.27	0.26	0.26
N	344148	344148	344148	344148	344148

Notes: Sample: Young people at state-schools with valid data on subject choices, gender, SES, and KS2 performance. Omitted categories are female and being in a gender-balanced school. Reporting regression coefficients. t statistics in parentheses. Stars indicate statistical significance: + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Cond. ρ reports intra-cluster correlation coefficient from reported model.

increase in the average academic selectivity of the subjects its pupils study.

Adding gender and school's gender balance to the model (M2) makes little difference to the previously described relationships between prior attainment and subject selectivity. Nevertheless, we see a statistically-significant association between an individual's gender and the academic selectivity of the subjects they study, with females on average studying a set of subjects that are 0.13 standard deviations more academically selective than their comparable male peers. School-level gender effects are particularly marked when comparing single-sex schools to gender balanced ones with both kinds of single-sex school entering pupils for more academically selective subjects.

Adding socioeconomic status to the model (M3) we find that both individuals' SES and the average SES in their school are positively associated with the academic selectivity of the subjects they study; these seem of similar qualitative importance. Furthermore, adding the controls for SES reduces the association between school's KS2 intake and subject choice markedly, although it remains an important factor.

Finally, in M4, we consider selective schools and those in local areas where academically selective schools take a significant proportion of individuals (specifically, local authority areas where at least 5% of the population attend a selective school). Including indicators for whether an individual is in a selective area and whether they are in a selective school in the model makes little difference to the other covariates. However, pupils in a school in a selective area study for a statistically significantly less academically selective mix of subjects. Those actually in selective schools see an approximately offsetting effect. This suggests that this result is driven by pupils in non-selective schools in selective areas studying less academically selective subjects, even after individual- and school-level factors have been taken into account.

Turning to our binary models of individuals studying particular combinations of subjects (Table 17), we restrict our discussion to highlighting important differences between the general subject selectivity model and these more specific models. We only report the models including the full set of covariates we consider (M4). We also report the intra-cluster correlation from this model (Cond. ρ) and from the model not including any covariates (Uncond. ρ) for comparison. For these models we report odds ratios (exponentiated logistic regression coefficients), which may be interpreted as the expected change in the odds of studying the relevant set of subjects.

We start our discussion with whether individuals study at least three 'facilitating' subjects, something that we find 81% of the sample to do. As robustness checks we have also run models of whether individuals study at least four or five 'facilitating' subjects and find broadly similar results at these alternative margins. The pattern of significant coefficients is broadly similar to the model of our

Table 17: Odds ratios from logistic regression models of probability that individuals study specific combinations of subjects at age 14

	3 Facil.	Applied	EBacc	Triple Sci.
KS2 Z-Score	2.022	0.861	2.116	1.060
	(70.76)***	(-12.19)***	(75.10)***	(27.81)***
School KS2 Z-Score	2.288	0.426	2.551	1.034
	(7.44)***	(-4.55)***	(8.14)***	(2.99)**
Male	0.969	0.710	0.918	1.016
	(-1.95)+	(-13.31)***	(-5.49)***	(13.93)***
All-male school	1.490	0.403	1.773	1.099
	(3.47)***	(-3.59)***	(4.53)***	(4.87)***
Mainly-male school	1.115	1.023	1.020	1.019
	(1.13)	(0.13)	(0.20)	(2.16)*
Mainly-female school	0.838	0.948	0.820	1.004
	(-2.58)**	(-0.44)	(-2.70)**	(0.69)
All-female school	2.398	0.477	2.063	0.988
	(7.65)***	(-3.71)***	(7.53)***	(-1.02)
SES Z-Score	1.279	0.993	1.253	1.004
	(36.00)***	(-1.02)	(31.44)***	(6.62)***
School SES Z-Score	1.209	0.857	1.581	1.001
	(3.29)***	(-1.59)	(6.95)***	(0.17)
>5% in LA in selective schools	0.824	1.077	0.761	0.987
	(-2.52)*	(0.55)	(-3.36)***	(-1.93)+
Selective school	3.909	0.0758	1.235	1.168
	(7.07)***	(-7.27)***	(1.26)	(6.09)***
Dep. Var. Mean	0.81	0.31	0.37	0.08
Uncond. ρ	0.40	0.67	0.42	0.29
Cond. ρ	0.30	0.64	0.32	0.23
N	344148	344148	344148	344148

Notes: Sample: Young people at state-schools with valid data on subject choices, gender, SES, and KS2 performance. Omitted categories are female and being in a gender-balanced school. Reporting odds ratios (i.e. exponentiated coefficients from the logistic regression model). t statistics in parentheses. Stars indicate statistical significance: + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Uncond. ρ reports intra-cluster correlation coefficient from model of dependent variable including no covariates; Cond. ρ reports intra-cluster correlation coefficient from reported model.

academic selectivity measure, although we cannot compare the magnitudes with the selectivity model due to its linear outcome. One standard deviation higher prior attainment by an individual at age 11 is associated with twice the odds of studying at least three facilitating subjects. Much less dramatically, individuals with one standard deviation higher SES have 28% increased odds of studying three or more facilitating subjects, while males have just slightly reduced odds of doing so.

There is a large and statistically significant association with school intake in terms of prior attainment: a one standard deviation increase in the intake of the school on this measure is associated with a more than doubling of the odds of studying at least three facilitating subjects. The overall importance of schools, as measured by the intra-cluster correlation, is similar to that in the overall model, both unconditionally and conditional on the set of controls. There is also evidence of increased odds of studying these subjects in single sex schools of either type, especially all-female schools. There is also an association with the SES intake of the school of a similar magnitude of that seen for the individual-level SES association. Even after accounting for these other characteristics, those in areas with selective schools see a significant reduction the odds of studying at least three facilitating subjects, while those actually in selective schools have almost four times the odds of doing so.

Whether individuals study any applied subjects, something that is true of 31% of the sample, has a contrasting set of associations. There is a negative relationship between individuals' prior attainment at age 11 and the odds of studying any applied subjects: individuals with a one standard deviation higher KS2 score have 14% lower odds of studying an applied subject. There is no significant difference between individual's SES and odds of studying an applied subject, although a negative relationship was observed in our survey data model, perhaps due to the stronger measurement of SES in that setting. One area of similarity between the overall selectivity model and the applied subjects model is that male students have reduced odds of studying any applied subjects. Likewise, individuals in all-male schools see one-quarter the odds of studying applied subjects compared to their peers in gender-balanced schools; a similar (but not quite as large) reduction is also evident in all-female schools. Again, the differences associated with selective schooling are stark: those in selective schools are over 90% lower odds of studying any applied subjects than peers in non-selective schools.

Another difference that stands out is the much greater value of intra-cluster correlation, compared to the general selectivity score model. Less than half of the variance is between individuals within the same schools, perhaps suggesting schools play a particularly large role in whether or not individuals study such subjects. Furthermore, this influence of schools seems divorced from the makeup of schools in this case, since the conditional intra-cluster correlation is only marginally lower than the unconditional one.

We next turn to the model of whether individuals are studying the full set of subjects necessary to be eligible for the English Baccalaureate (EBacc) if they also go on to reach the required standard in the qualifications. Just over a third (37%) of the sample also meet this criteria. Individuals with one standard deviation higher KS2 scores have over twice the odds of studying the EBacc subjects, while those in schools with one standard deviation higher KS2 intake see their odds of studying all EBacc subjects increase by over two and a half times. Individuals with one standard deviation higher SES have approximately 25% increased odds; those in schools with a one standard deviation higher intake in terms of SES see their odds increase by just over 50%. These are broadly similar patterns to those seen in the models of general subject selectivity in terms of the significance of individuals' prior attainment and SES, and schools' average intake in terms of these two characteristics.

The individual-level gender difference is much smaller than for the overall selectivity score but, once again, there is a much increased chance of taking EBacc subjects in all-male or all-female schools. Again, the presence of selective schooling in an area appears to be associated with a reduced probability of taking a full set of subjects, except in selective schools themselves where the opposite is true. Finally, schools also appear to play a bigger role in explaining whether young people study EBacc subjects than they did in explaining our general selectivity score, with an conditional intra-cluster correlation of 0.37.

We next consider variation in whether individuals study three separate sciences. This is the smallest group that we consider, at only 8% of the sample. Once again, those in selective schools are more likely to study triple sciences, even after taking into account other characteristics in the model. Students in an all male schools see 10% increased odds of studying this subject combination, relative to those in gender balanced schools while, in contrast to other subject combinations we have considered, there is no significant change in the odds among pupils in all-female schools. Even holding these influences of school constant, male pupils are a little more likely to study three separate sciences than their female peers. This is in contrast to findings for overall academic subject selectivity, where male pupils were more likely to study less academically selective subjects.

Conditional on the covariates in the model, and in contrast to the other subject combination models, whether individuals study triple sciences sees a lower proportion of variation explained by schools to that estimated in overall selectivity score, with a conditional intra-cluster correlation of 0.22.

6 Combinations of subject choices and university entry

This section borrows techniques from the programme evaluation literature to consider whether young people who study the full set of subjects required for EBacc-eligibility between ages 14 and 16 have different probabilities of applying to university, entering university and attending a high-status university. It also examines the same issue for sub-elements of the EBacc: studying for two or more science qualifications (either separate or combined awards), studying a foreign language, and studying History or Geography (we do not consider English or Maths since these are mandatory). By way of contrast, we also consider differences by whether individuals study for any 'applied' GCSEs, which, it has been argued, provide less effective preparation for future university study.

In particular, this paper contrasts purely descriptive differences in outcomes to those from flexible regression adjustment and matching approaches. These attempt to compare individuals very close to the margin of studying each full set of subjects, adjusting for observable differences in a highly flexible manner, taking advantage of rich survey data from a recent cohort of young people in England. As such, the estimates from this method demonstrate just the impact of having studied the full combination of subjects, rather than of the cumulative changes from overall differences in curriculum. To provide context for these results, we also produce estimates of the change in probability of the university outcomes associated with a continuous change in the academic selectivity of subjects studied and conditional on the same set of observable characteristics in the main analysis.

6.1 Methods

This section discusses our analytical strategy for comparing individuals' probabilities of going to university depending upon their subject choices, while taking into account that these individuals may well differ in other important respects. We take two main approaches, first applying binary choice regression modelling of our outcomes of interest, then using propensity score matching approaches to account more flexibly for differences in background characteristics. These methods both have advantages and disadvantages relative to one another, which we highlight in the following discussion.

6.1.1 Regression analysis

Regression modelling is a well-established method for estimating the association between a treatment variable and outcomes of interest, holding other background characteristics constant. Its advantage is that it provides an estimate of the treatment across the sample; however, this is also a disadvantage in that it may be extrapolating beyond the sample for which the data can provide us with reliable evidence. It also relies upon the regression equation adequately describing the relationship between independent and dependent variables.

Since our outcomes of interest are dichotomous, we estimate linear probability regression models (analyses using probit models do not give qualitatively different results). We use the following regression specification, recommended by Imbens and Rubin (2015), to estimate difference in outcomes conditional on a vector of controls, $\mathbf{X}$, listed below:

$$Y_i = \alpha + \beta \text{Treat}_i + \gamma(\mathbf{X}'_i - \bar{\mathbf{X}}') + \delta \text{Treat}_i(\mathbf{X}'_i - \bar{\mathbf{X}}') + \varepsilon_i \tag{4}$$

where Y is a binary indicator of whether individuals achieve our outcome of interest and Treat is a binary indicator of our subject combinations. In this regression, β is our primary coefficient of interest, recovering the average conditional difference in outcomes associated with the subject choice variable (separately: studying subjects required to be eligible for EBacc, studying at least one foreign language, studying two or more sciences, and studying for at least one applied GCSE).

This approach attempts to isolate the conditional association between subject choices and university access outcomes by using the extremely rich background data available in Next Steps. Specifically, we include the following covariates as dummy variables (where they are categorical) or linear and quadratic variables (where they are continuous): household income; age 14 (KS3) test scores (English, maths and science); gender; ethnic group (white, mixed, Indian, Pakistani, Black Caribbean, Black African, other); month of birth (continuous linear); number of siblings (categorical: none, one, two, or three or more); number of elder siblings (categorical: none, one, two, or three or more); lone parent family; mother's qualifications (none, below GCSEs, A Levels, HE below degree, degree); father's qualifications (none, below GCSEs, A Levels, HE below degree, degree); region of England (North East, North West, Yorkshire & Humber, East Midlands, West Midlands, East of England, South East, South West); school type (community, community technology college, foundation school, voluntary aided, voluntary controlled); whether school is selective; whether school has sixth form; mother's occupational status; and father's occupational status.

Given the design of Next Steps, we use clustered standard errors to account for the additional uncertainty around our estimates that this implies.

6.1.2 Matching

We also use propensity score matching methods (Rosenbaum and Rubin, 1983) to provide estimates of the conditional change in probability of university attendance. Mendolia and Walker (2014), Alcott (2017) and McDool (2017) have previously used this approach to address research questions using Next Steps data. It has the advantage of controlling for background characteristics in a more flexible manner. It also more explicitly restricts attention to the sample within which the data can provide reliable causal impacts (imposing 'common support'), rather than extrapolating across the sample.

However, to produce its causal estimates, it is important to stress that it still relies on the assumption of all differences between individuals in the ‘treated’ and ‘untreated’ groups being captured by observed characteristics included in the propensity score model. In this case, ‘treated’ corresponds to individuals who study the combination of subjects considered (separately: studying subjects required to be eligible for EBacc, studying at least one foreign language, studying two or more sciences, and studying for at least one applied GCSE). Next Steps’ rich set of such characteristics helps to make this a plausible assumption but we cannot, of course, rule out the continued presence of unobserved factors that determine the subject choices that individuals make (Shadish, 2012).

In this section, we lay out the matching approaches that we consider and discuss how we will assess whether they generate a matched sample that is well balanced on our observable characteristics and has good common support. The subsequent section reports on the process of constructing matched samples and assessing how well balanced these are in terms of background variables.

Matching begins by specifying a model of whether individuals study each of these sets of subjects starting at age 14. These are discrete choice models, specifically in this case we use a probit link function. This model includes the same set of background characteristics as those added to the regression model discussed in Section 6.1.1. However, we also experimented with additional complexity in the model, such as the inclusion of interaction terms between characteristics, where this helped to increase the balance.

This model is used to generate an estimated propensity score for each individual i.e. the estimated probability that they study the relevant combination of subjects (are ‘treated’ in the policy evaluation terminology). We consider the distribution of these estimated propensity scores among treated and untreated individuals in order to assess the extent to which they overlap and which matching approaches are likely to construct a balanced sample.

These estimated propensity scores are then used to produce a matched sample. We consider three methods of doing this:

- Nearest neighbour matching without replacement, with caliper: each treated individual is matched to one untreated individual with the closest propensity score, subject to the constraint (caliper) that the scores are no more than 0.05 different; once an individual has been used as a match they cannot be used again.
- Nearest neighbour matching with replacement, with caliper: each treated individual is matched to one untreated individual with the closest propensity score, subject to the constraint (caliper) that the scores are no more than 0.05 different; an individual can be used as a match multiple

times.

- Kernel matching: each treated individual is matched to all untreated individuals with a weighting scheme that gives closer matches larger weight.

We assess the matched samples produced by these methods by considering the standardised differences in the background characteristics included in the matching model. Standardised differences are constructed by dividing the absolute difference in the characteristic between the treatment and control groups by the overall standard deviation of the characteristics, meaning that they are all in a common scale. As well as considering the average standardised difference across all characteristics of the matched sample, we also consider each characteristic to ensure that all differences are acceptably small (Imbens and Rubin, 2015).

Finally, we estimate linear probability regression models of our outcomes of interest using the matched sample, using the same model as that described in section 6.1.1.

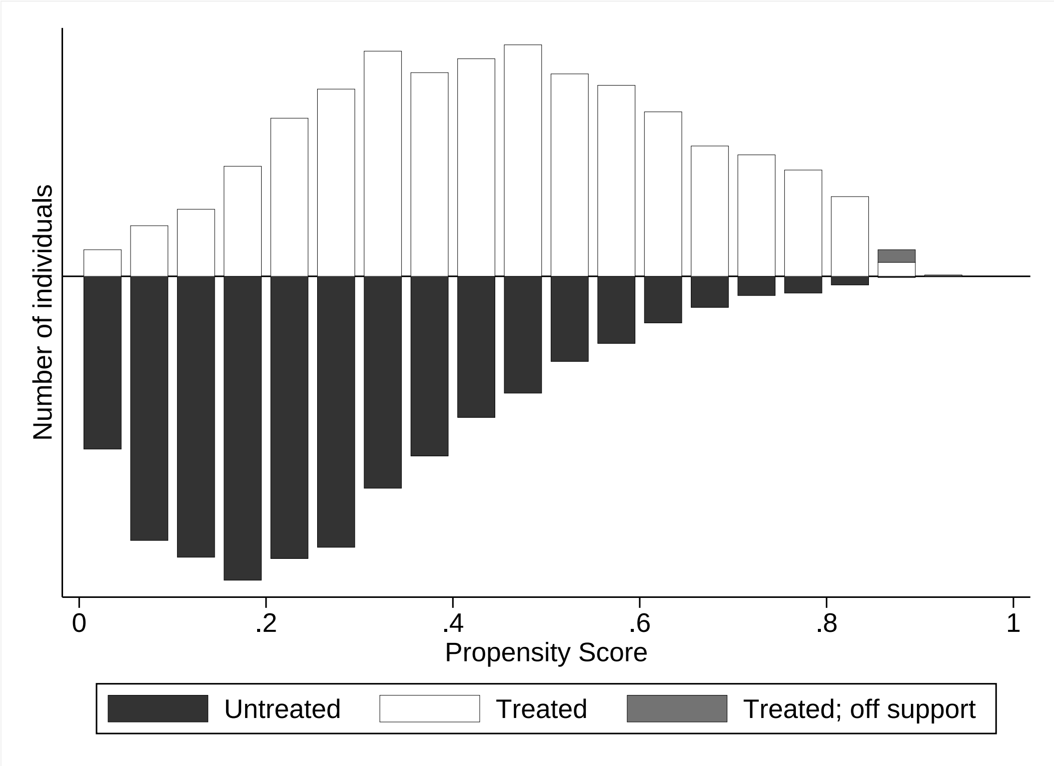
6.2 Constructing a matched sample

The distribution of the propensity scores by combination of subjects studied is shown in Figure 7 for whether individuals studied the subjects required to be eligible for EBacc, Figure 8 for whether individuals studied two or more sciences, Figure 9 for whether individuals studied foreign languages, Figure 10 for whether individuals studied either History or Geography, and Figure 11 for whether individuals studied any applied GCSEs. The distributions have implications for the preferred matching approach to take, which are now discussed in turn.

The distribution of the propensity score by whether individuals study a full set of subjects required to be eligible for EBacc (Figure 7) is overlapping across most of its range. However, above propensity scores of approximately 0.3 there are far fewer untreated individuals than there are treated individuals. While this is unsurprising, our approach to matching will need to ensure that treated individuals with high propensity scores do not end up being matched to untreated individuals with much lower propensity scores due to a lack of available matches of a similar score. Possible solutions include imposing a caliper width on suitable matches or allowing matching with replacement (i.e. more than one treated individual is matched to the same untreated individual), both of which will reduce the size of the matched sample. There are only a small number of treated individuals whose propensity score is higher than that of the highest propensity score among untreated individuals who are, therefore, considered outside the range of common support (referred to as being ‘off support’) and excluded from the matched sample.

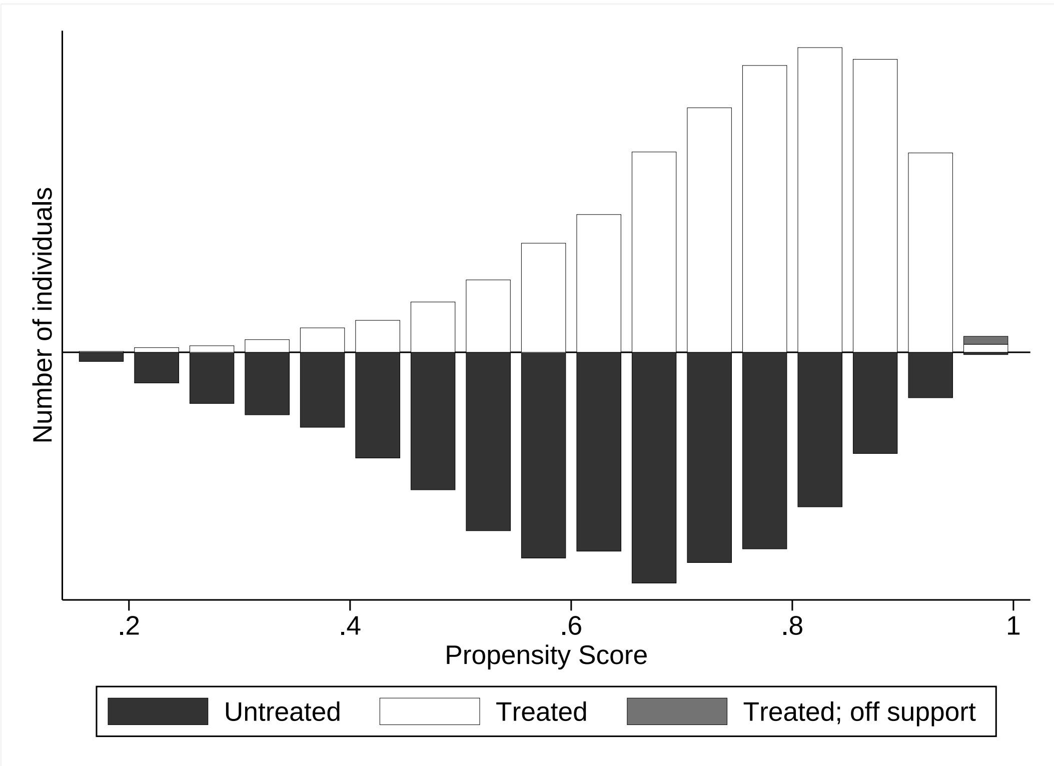
The distribution of the propensity score by whether individuals study two or more sciences (or for a

Figure 7: Distribution of propensity scores by whether individuals would be eligible for EBacc



Notes: Histogram of the propensity score for treated individuals (i.e. study all subjects required to be eligible for EBacc) above x axis and untreated individuals (i.e. do not study all subjects required to be eligible for EBacc) below x axis. Individuals described as 'off support' fall outside the range of common support and so are excluded from impact estimation.

Figure 8: Distribution of propensity scores by whether individuals study two or more sciences



Notes: Histogram of the propensity score for treated individuals (i.e. study two or more science subjects) above x axis and untreated individuals (i.e. do not study two or more science subjects) below x axis. Individuals described as 'off support' fall outside the range of common support and so are excluded from impact estimation.

double award) is more skewed towards 1 compared to the distribution for EBacc-eligibility, reflecting the larger proportion of the sample (67%) who meet this criterion. Once again, there are only a small number of treated individuals whose propensity score is higher than that of the highest propensity score among untreated individuals who are, therefore, considered outside the range of common support ('off support') and excluded from the matched sample.

Turning to propensity scores defined by whether or not individuals study any foreign languages, the distributions are more obviously overlapping than was the case for the EBacc sample. Nevertheless, at the tails of the distribution there is mismatch in the relative number of treated and untreated individuals, especially at high propensity scores. A large number of treated individuals have propensity scores very close to 1 and, as a result, a more substantial number of treated individuals are excluded due to being 'off support'.

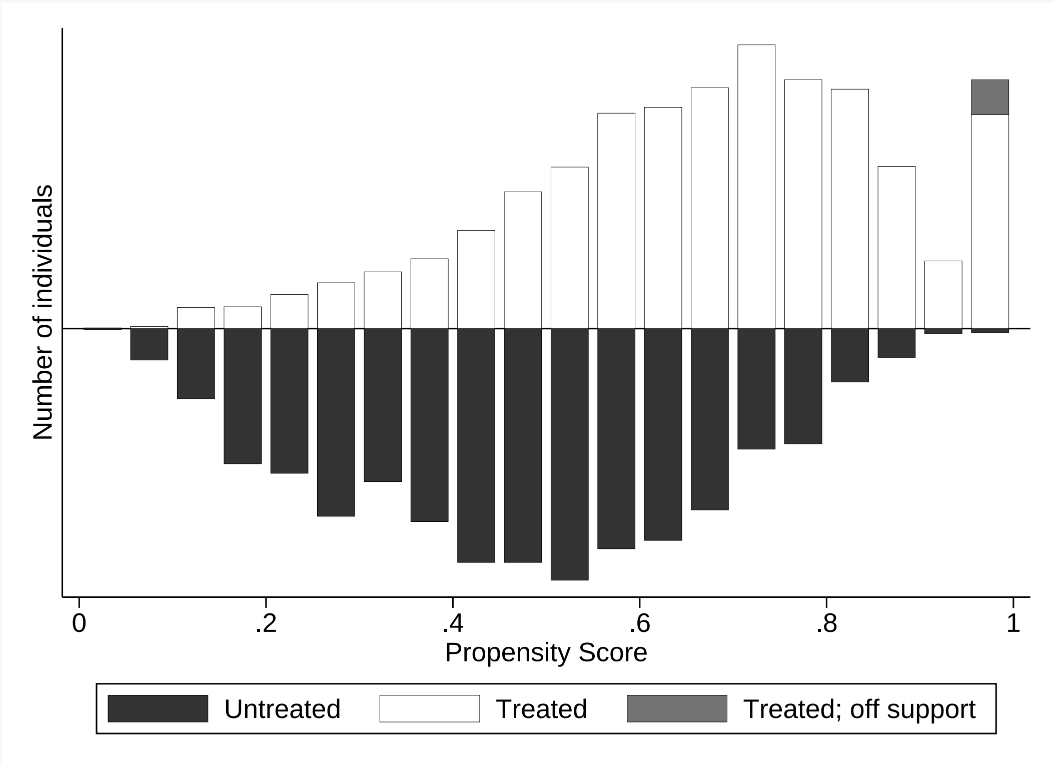
Next, we consider the distribution of propensity scores by whether individuals study either History, Geography or both. There is a strong overlap across almost the entire range of the propensity score distribution with only very few individuals who study History or Geography 'off support' (outside the range of common support) due to having a higher estimated propensity than the highest score estimated for individuals who do not study either of these subjects.

Finally, we consider the propensity score distribution defined by whether or not individuals study for any applied GCSEs. Again, there is a good overlap across much of the range. There are no treated individuals with propensity scores falling outside the range of common support.

On the basis of these distributions, our preferred method of matching is likely to be a nearest neighbour matching approach, without replacement (i.e. once an untreated individual is selected as a match they cannot be selected again) imposing common support and a caliper of 0.05. As discussed, imposing common support does not result in the exclusion of many observations since there is a similar range of propensity scores in treatment and control groups but remains important to exclude treated individuals for whom there are no comparable untreated individuals. There are also large parts of the propensity score distribution of treated individuals in which there are few suitable untreated matches. As such, a caliper on the distance between the propensity score of the treated individual and that of the match ensures that untreated matches do not end up being too different from their treated comparator.

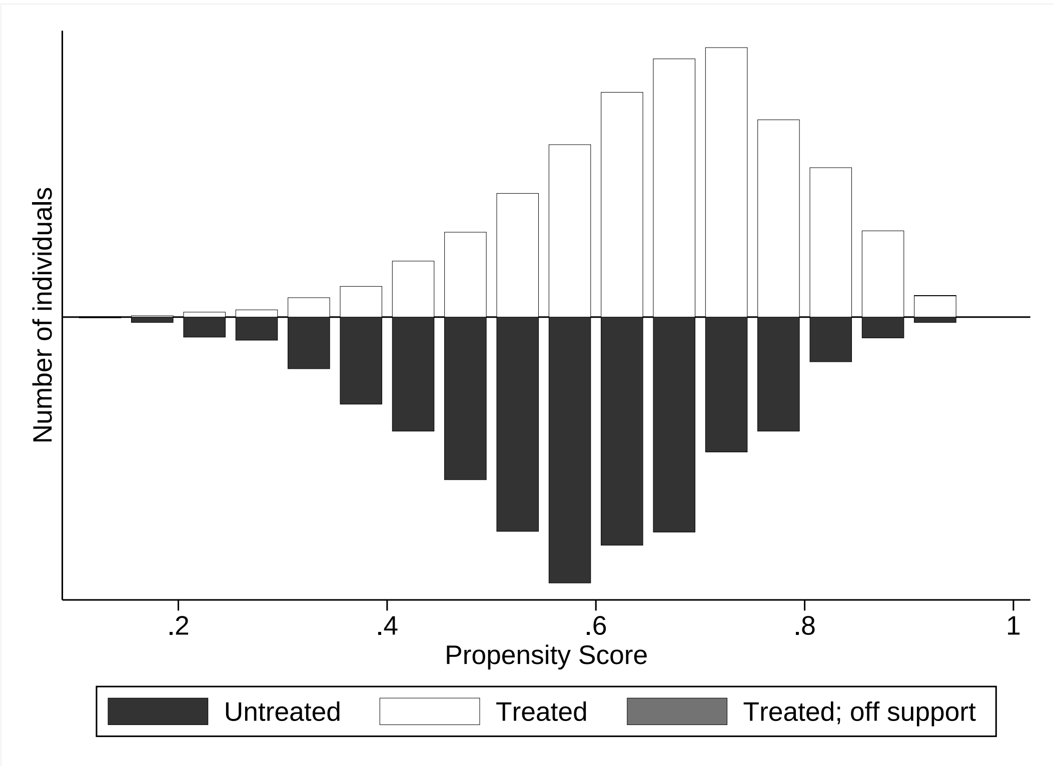
Nevertheless, we do also perform our other proposed forms of matching. Matched samples from these approaches do, in some cases, differ somewhat from our preferred approach. Matching with replacement resulted in significantly smaller sample sizes than from applying a caliper. This suggests that the results are relying on a small number of untreated individuals being used as matches many

Figure 9: Distribution of propensity scores by whether individuals study any foreign languages



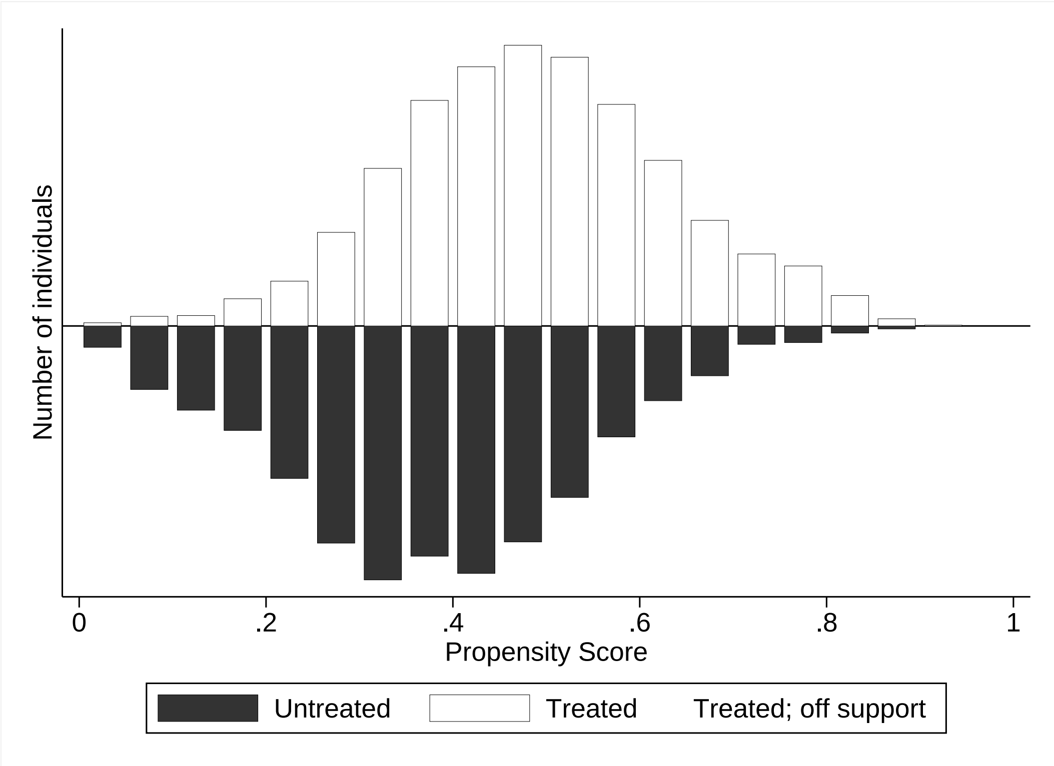
Notes: Histogram of the propensity score for treated individuals (i.e. study at least one foreign language) above x axis and untreated individuals (i.e. do not study at least one foreign language) below x axis. Individuals described as 'off support' fall outside the range of common support and so are excluded from impact estimation.

Figure 10: Distribution of propensity scores by whether individuals study History or Geography



Notes: Histogram of the propensity score for treated individuals (i.e. study History, Geography or both) above x axis and untreated individuals (i.e. do not study either History or Geography) below x axis. Individuals described as 'off support' fall outside the range of common support and so are excluded from impact estimation.

Figure 11: Distribution of propensity scores by whether individual study for any applied GCSEs



Notes: Histogram of the propensity score for treated individuals (i.e. study for at least one applied GCSE) above x axis and untreated individuals (i.e. do not study for any applied GCSEs) below x axis. Individuals described as 'off support' fall outside the range of common support and so are excluded from impact estimation.

times, which is not robust and is more likely to result in less representative results. In addition, the overall reduction in bias (as measured by standardised differences) was smaller, sometimes considerably, in matching with replacement. Kernel matching also resulted in smaller reductions in bias than nearest-neighbour matching with a caliper.

Since each individual in the treatment group is matched to an untreated individual with as similar as possible a propensity score, the matched sample should be balanced on the characteristics included in the propensity score model. We verify that this is the case in Table 18 for our matched sample by EBacc-eligibility, Table 19 for our matched sample by studying any languages, Table 20 for our matched sample by two or more sciences, Table 21 for our matched sample by whether individuals study History or Geography, and Table 22 for our matched sample by applied subjects.

Table 18 suggests that matching has produced a strongly balanced sample by whether or not individuals study a full set of EBacc subjects. While, in the unmatched sample, there are substantial standardised differences between variables such as household income, prior attainment and parental education, these have all been substantially reduced in the matched sample. No standardised differences exceed 0.08 in the matched sample. A small number of characteristics see an increased bias (negative figures in the % bias reduction column), but this is driven by only small increases in very

Table 18: Balance of characteristics by whether individuals would be eligible for EBacc

Variable	Unmatched Std. Diff	Matched Std. Diff	% Bias Red.	p
KS3 English Z-Score	0.75	0.00	99.9	0.98
KS3 Maths Z-Score	0.76	0.00	99.8	0.97
KS3 Science Z-Score	0.80	0.01	98.7	0.79
H/h income (£)	0.48	0.02	95.5	0.53
Male	0.03	0.00	95.6	0.97
Ethnic group: Mixed	0.03	0.00	88.1	0.93
Ethnic group: Indian	0.03	0.01	54.7	0.77
Ethnic group: Pakistani	0.05	0.02	59.4	0.58
Ethnic group: Bangladeshi	0.13	0.00	97.1	0.93
Ethnic group: Black Caribbean	0.01	0.00	42.4	0.89
Ethnic group: Black African	0.03	0.02	30.8	0.61
Ethnic group: Other	0.02	0.02	37.0	0.65
No Siblings	0.01	0.00	66.2	0.91
2 Siblings	0.02	0.03	-27.2	0.45
3+ Siblings	0.17	0.03	85.4	0.52
No Older Siblings	0.11	0.02	83.6	0.62
2 Older Siblings	0.07	0.01	87.4	0.79
3+ Older Siblings	0.11	0.02	85.5	0.64
Lone parent family	0.08	0.02	79.5	0.67
Mother has no quals.	0.25	0.02	92.4	0.60
Mother has below GCSEs	0.13	0.01	96.0	0.89
Mother has A Levels	0.06	0.02	67.4	0.57
Father has HE below degree	0.16	0.00	98.9	0.96
Mother has degree	0.43	0.04	91.4	0.27
Father has no quals.	0.23	0.01	97.2	0.86
Father has below GCSEs	0.13	0.01	93.3	0.81
Father has A Levels	0.03	0.03	11.5	0.40
Father has HE below degree	0.10	0.00	96.4	0.92
Father has degree	0.48	0.04	90.7	0.23
Region: North East	0.01	0.01	-15.0	0.88
Region: North West	0.00	0.01	-248.0	0.81
Region: Yorkshire & Humber	0.06	0.00	100.0	1.00
Region: East Midlands	0.02	0.02	3.3	0.70
Region: West Midlands	0.04	0.03	26.5	0.55
Region: East of England	0.07	0.01	78.9	0.77
Region: South East	0.13	0.03	79.6	0.64
Region: South West	0.01	0.00	100.0	1.00
School: Community Tech. College	0.01	0.01	44.1	0.92
School: Foundation School	0.14	0.01	91.9	0.82
School: Voluntary Aided	0.11	0.03	72.7	0.55
School: Voluntary Controlled	0.07	0.01	90.4	0.90
Selective School	0.71	0.08	88.8	0.07
School w/ 6th Form	0.20	0.01	93.4	0.80
Mother: Long-term unemployed	0.21	0.03	85.3	0.45
Mother: Intermediate Occupations	0.10	0.00	98.5	0.97
Mother: Higher Occupations	0.31	0.00	99.2	0.94
Father: Long-term unemployed	0.10	0.03	71.9	0.46
Father: Intermediate Occupations	0.04	0.02	60.4	0.63
Father: Higher Occupations	0.45	0.01	98.1	0.82
Overall	0.17	0.02	88.2	

Notes: Std. Diff = Standardised difference (absolute difference between two groups divided by standard deviation of pooled sample); p = p value of a t test of the null hypothesis of no difference between the treated and untreated groups in the matched sample.

small pre-matching standardised differences. Overall, the average standardised difference between characteristics in the treatment and control group are reduced from 0.17 to 0.02.

The sample matched on whether individuals study any foreign languages (Table 19) also sees much improvement in terms of balance. Particularly large differences are evident in the unmatched sample (with an average standardised difference of 0.22), especially for household income, parental education and school type; these are much reduced in the matched sample. A small number of characteristics are slight outliers, although still with modest standardised differences of 0.07-0.08. Overall, as with the matched sample for EBacc, the average standardised difference between treatment and control groups is dramatically reduced to 0.03.

A similar picture is also evident for the sample defined by whether or not individuals study two or more sciences. There are, again, a couple of outliers in terms of standardised differences, with selective school attendance and ethnic group Indian having standardised differences of 0.08 and 0.09, respectively, in the matched sample; in this case these are not statistically significant differences. Overall, the matching exercise results in a 75% reduction in bias on observable characteristics.

The only outlier in terms of balance, in the case of the matched sample divided by whether individuals study History or Geography, is whether mother has a degree. This has a standardised difference of 0.07 and this is still not statistically significant. The overall balance is reduced from an average standardised difference of 0.12 in the unmatched sample to an average standardised difference of 0.02 in the matched sample.

Finally, we see a similarly well-matched sample when splitting the group by whether they study for any applied GCSEs. In the matched dataset the standardised differences between groups does not exceed 0.04 for any of the characteristics. The average standardised difference in the matched dataset is 0.01, compared to 0.12 in the unmatched sample.

6.3 Results

The results are reported in Table 23. This includes:

1. the 'naïve' estimates of the change in probability of university access measures associated with studying the relevant set of subjects (these replicate the difference between the two columns in the top panel of Table 25);

2. the regression adjusted estimates of the change in probability of university access measures having controlled for background characteristics parametrically;

3. and, finally, the matched estimates of the change in probability of university access measures, by estimating conditional differences among those in the matched samples constructed in Sec-

Table 19: Balance of characteristics by whether individuals are studying any foreign languages

Variable	Unmatched Std. Diff	Matched Std. Diff	% Bias Red.	p
KS3 English Z-Score	0.80	0.07	91.3	0.09
KS3 Maths Z-Score	0.75	0.04	94.7	0.31
KS3 Science Z-Score	0.75	0.05	93.2	0.20
H/h income (£)	0.53	0.07	86.0	0.03
Male	0.18	0.07	58.9	0.05
Ethnic group: Mixed	0.03	0.00	91.2	0.93
Ethnic group: Indian	0.02	0.02	8.9	0.66
Ethnic group: Pakistani	0.15	0.00	98.5	0.96
Ethnic group: Bangladeshi	0.20	0.02	88.9	0.60
Ethnic group: Black Caribbean	0.03	0.00	85.7	0.89
Ethnic group: Black African	0.00	0.00	100.0	1.00
Ethnic group: Other	0.03	0.02	30.2	0.52
No Siblings	0.05	0.00	100.0	1.00
2 Siblings	0.01	0.01	15.0	0.80
3+ Siblings	0.25	0.00	98.9	0.94
No Older Siblings	0.14	0.01	90.2	0.69
2 Older Siblings	0.05	0.01	75.8	0.75
3+ Older Siblings	0.18	0.01	94.7	0.81
Lone parent family	0.03	0.00	88.5	0.92
Mother has no quals.	0.32	0.03	91.9	0.53
Mother has below GCSEs	0.11	0.04	64.1	0.28
Mother has A Levels	0.13	0.01	88.5	0.66
Father has HE below degree	0.19	0.02	89.2	0.56
Mother has degree	0.40	0.07	82.2	0.04
Father has no quals.	0.30	0.02	93.1	0.60
Father has below GCSEs	0.08	0.01	92.0	0.85
Father has A Levels	0.06	0.01	80.6	0.74
Father has HE below degree	0.14	0.02	88.2	0.63
Father has degree	0.41	0.08	81.3	0.02
Region: North East	0.08	0.02	77.4	0.80
Region: North West	0.01	0.02	-93.4	0.77
Region: Yorkshire & Humber	0.10	0.00	98.2	0.98
Region: East Midlands	0.07	0.00	96.9	0.97
Region: West Midlands	0.01	0.01	24.8	0.87
Region: East of England	0.03	0.02	51.6	0.79
Region: South East	0.15	0.01	95.6	0.92
Region: South West	0.03	0.02	43.0	0.79
School: Community Tech. College	0.05	0.02	60.6	0.60
School: Foundation School	0.14	0.04	70.6	0.52
School: Voluntary Aided	0.16	0.02	89.6	0.81
School: Voluntary Controlled	0.06	0.01	84.1	0.88
Selective School	2.15	0.04	98.0	0.04
School w/ 6th Form	0.18	0.02	89.6	0.77
Mother: Long-term unemployed	0.26	0.01	98.0	0.91
Mother: Intermediate Occupations	0.14	0.04	70.4	0.24
Mother: Higher Occupations	0.35	0.07	81.0	0.07
Father: Long-term unemployed	0.16	0.03	83.3	0.46
Father: Intermediate Occupations	0.08	0.05	36.1	0.11
Father: Higher Occupations	0.47	0.08	81.9	0.02
Overall	0.22	0.03	86.4	

Notes: Std. Diff = Standardised difference (absolute difference between two groups divided by standard deviation of pooled sample); p = p value of a t test of the null hypothesis of no difference between the treated and untreated groups in the matched sample.

Table 20: Balance of characteristics by whether individuals are studying two or more sciences

Variable	**Unmatched Std. Diff**	**Matched Std. Diff**	**% Bias Red.**	**p**
KS3 English Z-Score	0.58	0.03	94.4	0.46
KS3 Maths Z-Score	0.63	0.00	99.4	0.93
KS3 Science Z-Score	0.68	0.02	96.8	0.61
H/h income (£)	0.34	0.04	86.9	0.49
Male	0.04	0.04	11.0	0.41
Ethnic group: Mixed	0.03	0.00	87.4	0.93
Ethnic group: Indian	0.08	0.09	-10.7	0.17
Ethnic group: Pakistani	0.02	0.01	42.2	0.78
Ethnic group: Bangladeshi	0.05	0.02	52.0	0.48
Ethnic group: Black Caribbean	0.03	0.05	-47.7	0.47
Ethnic group: Black African	0.03	0.04	-33.1	0.45
Ethnic group: Other	0.00	0.00	-3.5	0.91
No Siblings	0.01	0.03	-484.2	0.49
2 Siblings	0.04	0.03	25.5	0.55
3+ Siblings	0.11	0.00	97.9	0.95
No Older Siblings	0.05	0.00	90.9	0.91
2 Older Siblings	0.04	0.02	42.7	0.54
3+ Older Siblings	0.09	0.00	99.1	0.98
Lone parent family	0.07	0.00	96.0	0.92
Mother has no quals.	0.16	0.05	70.1	0.28
Mother has below GCSEs	0.12	0.03	76.6	0.38
Mother has A Levels	0.07	0.04	45.4	0.40
Father has HE below degree	0.10	0.03	74.4	0.61
Mother has degree	0.33	0.05	86.1	0.55
Father has no quals.	0.15	0.02	83.8	0.55
Father has below GCSEs	0.08	0.03	62.7	0.41
Father has A Levels	0.03	0.03	-22.4	0.46
Father has HE below degree	0.08	0.00	97.2	0.96
Father has degree	0.33	0.00	99.2	0.98
Region: North East	0.02	0.01	67.9	0.91
Region: North West	0.04	0.05	-29.3	0.39
Region: Yorkshire & Humber	0.11	0.01	90.8	0.82
Region: East Midlands	0.01	0.04	-303.5	0.61
Region: West Midlands	0.01	0.00	86.6	0.98
Region: East of England	0.02	0.03	-33.0	0.50
Region: South East	0.03	0.04	-31.0	0.39
Region: South West	0.01	0.04	-484.0	0.25
School: Community Tech. College	0.05	0.00	90.8	0.88
School: Foundation School	0.09	0.03	67.4	0.61
School: Voluntary Aided	0.09	0.02	77.2	0.70
School: Voluntary Controlled	0.03	0.01	56.8	0.85
Selective School	0.31	0.08	74.5	0.54
School w/ 6th Form	0.06	0.00	96.1	0.96
Mother: Long-term unemployed	0.11	0.02	79.5	0.56
Mother: Intermediate Occupations	0.07	0.00	96.5	0.96
Mother: Higher Occupations	0.20	0.02	87.7	0.67
Father: Long-term unemployed	0.08	0.02	72.9	0.58
Father: Intermediate Occupations	0.01	0.05	-281.2	0.21
Father: Higher Occupations	0.32	0.01	95.8	0.80
Overall	0.12	0.03	75.0	

Notes: Std. Diff = Standardised difference (absolute difference between two groups divided by standard deviation of pooled sample); p = p value of a t test of the null hypothesis of no difference between the treated and untreated groups in the matched sample.

Table 21: Balance of characteristics by whether individuals study History or Geography

Variable	Unmatched Std. Diff	Matched Std. Diff	% Bias Red.	p
KS3 English Z-Score	0.46	0.01	98.8	0.88
KS3 Maths Z-Score	0.46	0.01	97.2	0.73
KS3 Science Z-Score	0.50	0.01	98.5	0.84
H/h income (£)	0.33	0.05	85.1	0.38
Male	0.10	0.04	60.0	0.31
Ethnic group: Mixed	0.03	0.02	39.7	0.61
Ethnic group: Indian	0.08	0.01	85.3	0.77
Ethnic group: Pakistani	0.07	0.01	79.9	0.69
Ethnic group: Bangladeshi	0.11	0.01	89.6	0.64
Ethnic group: Black Caribbean	0.02	0.03	-61.3	0.36
Ethnic group: Black African	0.05	0.01	89.3	0.85
Ethnic group: Other	0.01	0.02	-103.6	0.65
No Siblings	0.02	0.02	17.8	0.65
2 Siblings	0.04	0.00	89.0	0.92
3+ Siblings	0.14	0.02	88.5	0.61
No Older Siblings	0.08	0.00	95.3	0.92
2 Older Siblings	0.04	0.02	42.1	0.51
3+ Older Siblings	0.12	0.01	90.3	0.69
Lone parent family	0.04	0.01	70.5	0.72
Mother has no quals.	0.22	0.00	98.4	0.91
Mother has below GCSEs	0.09	0.02	83.7	0.60
Mother has A Levels	0.01	0.01	-148.4	0.71
Father has HE below degree	0.09	0.00	97.1	0.96
Mother has degree	0.29	0.07	76.0	0.29
Father has no quals.	0.20	0.03	85.6	0.36
Father has below GCSEs	0.06	0.01	84.3	0.78
Father has A Levels	0.05	0.01	79.0	0.80
Father has HE below degree	0.03	0.01	82.7	0.89
Father has degree	0.33	0.02	93.2	0.73
Region: North East	0.01	0.01	5.5	0.87
Region: North West	0.02	0.03	-52.4	0.53
Region: Yorkshire & Humber	0.00	0.00	30.7	0.97
Region: East Midlands	0.03	0.00	88.0	0.94
Region: West Midlands	0.07	0.01	83.5	0.77
Region: East of England	0.03	0.02	47.2	0.75
Region: South East	0.09	0.01	93.2	0.91
Region: South West	0.03	0.03	-2.1	0.46
School: Community Tech. College	0.01	0.01	27.7	0.82
School: Foundation School	0.14	0.00	98.4	0.97
School: Voluntary Aided	0.07	0.02	71.0	0.74
School: Voluntary Controlled	0.11	0.04	64.0	0.62
Selective School	0.31	0.01	96.9	0.92
School w/ 6th Form	0.14	0.02	88.6	0.73
Mother: Long-term unemployed	0.18	0.01	95.6	0.77
Mother: Intermediate Occupations	0.09	0.00	96.5	0.94
Mother: Higher Occupations	0.19	0.04	80.3	0.39
Father: Long-term unemployed	0.09	0.03	70.0	0.43
Father: Intermediate Occupations	0.03	0.02	39.0	0.60
Father: Higher Occupations	0.30	0.03	89.6	0.48
Overall	0.12	0.02	83.3	

Notes: Std. Diff = Standardised difference (absolute difference between two groups divided by standard deviation of pooled sample); p = p value of a t test of the null hypothesis of no difference between the treated and untreated groups in the matched sample.

Table 22: Balance of characteristics by whether individuals are studying for any applied GCSEs

Variable	**Unmatched Std. Diff**	**Matched Std. Diff**	**% Bias Red.**	**p**
KS3 English Z-Score	0.52	0.04	93.0	0.30
KS3 Maths Z-Score	0.51	0.03	93.2	0.32
KS3 Science Z-Score	0.54	0.04	92.1	0.22
H/h income (£)	0.32	0.02	94.3	0.61
Male	0.02	0.01	71.7	0.87
Ethnic group: Mixed	0.07	0.01	92.5	0.86
Ethnic group: Indian	0.03	0.01	62.8	0.75
Ethnic group: Pakistani	0.09	0.00	100.0	1.00
Ethnic group: Bangladeshi	0.13	0.02	88.2	0.69
Ethnic group: Black Caribbean	0.02	0.00	76.9	0.91
Ethnic group: Black African	0.01	0.00	69.3	0.92
Ethnic group: Other	0.05	0.02	47.9	0.49
No Siblings	0.01	0.01	-15.2	0.64
2 Siblings	0.00	0.00	55.9	0.94
3+ Siblings	0.13	0.01	95.3	0.85
No Older Siblings	0.04	0.01	78.0	0.79
2 Older Siblings	0.03	0.00	100.0	1.00
3+ Older Siblings	0.10	0.03	69.7	0.32
Lone parent family	0.03	0.01	73.0	0.75
Mother has no quals.	0.23	0.01	94.7	0.73
Mother has below GCSEs	0.15	0.04	74.4	0.17
Mother has A Levels	0.05	0.00	97.6	0.97
Father has HE below degree	0.11	0.00	96.2	0.89
Mother has degree	0.23	0.04	83.5	0.25
Father has no quals.	0.19	0.01	97.1	0.87
Father has below GCSEs	0.07	0.01	83.5	0.73
Father has A Levels	0.01	0.01	-62.5	0.79
Father has HE below degree	0.07	0.00	100.0	1.00
Father has degree	0.24	0.04	82.4	0.20
Region: North East	0.07	0.00	97.0	0.96
Region: North West	0.02	0.00	100.0	1.00
Region: Yorkshire & Humber	0.10	0.01	86.5	0.77
Region: East Midlands	0.03	0.01	70.4	0.81
Region: West Midlands	0.10	0.00	100.0	1.00
Region: East of England	0.01	0.02	-124.1	0.65
Region: South East	0.10	0.02	75.5	0.60
Region: South West	0.01	0.00	54.1	0.94
School: Community Tech. College	0.16	0.02	85.3	0.13
School: Foundation School	0.12	0.01	92.4	0.86
School: Voluntary Aided	0.12	0.04	66.4	0.38
School: Voluntary Controlled	0.05	0.00	94.3	0.94
Selective School	0.26	0.01	97.3	0.87
School w/ 6th Form	0.03	0.00	84.0	0.91
Mother: Long-term unemployed	0.16	0.02	89.4	0.64
Mother: Intermediate Occupations	0.07	0.01	92.6	0.87
Mother: Higher Occupations	0.19	0.01	93.5	0.70
Father: Long-term unemployed	0.05	0.00	95.3	0.94
Father: Intermediate Occupations	0.01	0.01	22.1	0.82
Father: Higher Occupations	0.29	0.02	93.7	0.58
Overall	0.12	0.01	91.7	

Notes: Std. Diff = Standardised difference (absolute difference between two groups divided by standard deviation of pooled sample); p = p value of a t test of the null hypothesis of no difference between the treated and untreated groups in the matched sample.

tion 6.2.

It also reports the statistical significance (taking into account the school-level clustering in Next Steps) of these estimated differences and the size of the matched sample. We report average marginal effects from linear probability models (average marginal effects from probit regression models give similar results).

As noted in Section 3, in the unmatched sample with no adjustments, individuals who study the full set of subjects requried to be eligible for the EBacc are 27 percentage points more likely to apply to university, and 29 percentage points more likely to attend, than their peers who do not study the full set of subjects. Once we control for background characteristics using regression analysis, these differences are dramatically reduced to differences of four and three percentage points, respectively. The differences remain statistically significant. In the matched sample, the difference in the probability of attending university remains statistically significant, however the difference in probability of applying does not. Surprisingly, given the particular rhetoric around EBacc representing the subjects favoured by more prestigious universities (Gibb, 2011), the results from the regression model imply that those with a full set of EBacc subjects are less likely to get into a Russell Group university than their peers who do not. However, this is not robust to using the matching approach, where the difference is essentially zero. This reduces our confidence in the regression adjustment result, suggesting it may be driven by extrapolation across not truly comparable individuals.

In purely descriptive terms, individuals who study two or more sciences are 24 percentage points more likely to apply to university and 25 percentage points more likely to attend than their peers who do not. As with the overall EBacc, these are much reduced once we adjust for background characteristics using the regression model but do remain statistically significant. However, they are reduced further and become statistically insignificant in the matched sample. Individuals who study two or more sciences are also 10 percentage points more likely to attend a Russell Group institution than those who only study for one science award. This is reduced to just a one percentage point difference in both the regression model and the matched sample; this is a statistically significant difference only when using the matching approach.

Students who study a foreign language are 25 percentage points more likely to apply and 26 percentage points more likely to attend university. However, once the background of those studying these subjects is taken into account, these differences are reduced to be small and statistically insignificant, whether we use regression modelling or consider our matched sample. In the case of entry to a Russell Group university, there is a difference between the regression adjustment and matching results. While those who study any languages are 11 percentage points more likely to attend a Russell Group university, the regression adjustment results suggest that, taking into account background

Table 23: Estimates of differences in university progression outcomes by combinations of subjects studied

Outcome	Unmatched	Regression	Matched
EBacc-eligible			
Apply to University	0.27	0.04	0.03
	(18.47)	(2.59)	(1.59)
Attend University	0.29	0.03	0.05
	(18.56)	(2.41)	(2.65)
Attend Russell Group	0.14	-0.02	0.01
	(10.34)	(-2.29)	(0.92)
N	6000	6000	2693
2+ Sciences			
Apply to University	0.24	0.04	0.02
	(14.75)	(3.14)	(1.27)
Attend University	0.25	0.05	0.02
	(16.20)	(3.78)	(1.20)
Attend Russell Group	0.10	0.01	0.01
	(10.87)	(1.44)	(2.39)
N	6000	6000	2758
Foreign Languages			
Apply to University	0.25	0.01	0.01
	(14.59)	(1.06)	(0.61)
Attend University	0.26	0.02	0.01
	(15.64)	(1.18)	(0.37)
Attend Russell Group	0.11	-0.04	0.00
	(10.63)	(-3.59)	(0.01)
N	6000	6000	3212
History or Geography			
Apply to University	0.17	0.03	0.02
	(10.25)	(2.03)	(1.24)
Attend University	0.16	0.02	-0.00
	(10.06)	(1.66)	(-0.07)
Attend Russell Group	0.08	0.02	-0.01
	(8.01)	(2.15)	(-1.01)
N	6000	6000	3753
Applied GCSEs			
Apply to University	-0.19	-0.04	-0.04
	(-12.03)	(-3.29)	(-2.73)
Attend University	-0.20	-0.04	-0.03
	(-12.63)	(-3.27)	(-2.16)
Attend Russell Group	-0.10	-0.02	-0.01
	(-9.26)	(-2.83)	(-1.11)
N	6000	6000	3958

Notes: Regression models are weighted using Next Steps-provided design and attrition weights. Figures in parentheses report t-statistics from test of null hypothesis of no difference, based on standard errors calculated taking into account school-level clustering. N in Matched analysis is purely for the treated (minus a small number removed due to being outside the range of common support) and matched controls, hence the reduced sample size.

characteristics, they are 4 percentage points less likely to do so. By contrast, the results from the matched sample suggest there is no statistically significant difference in Russell Group attendance by whether or not individuals have studied any languages.

Students who study either History or Geography are 17 percentage points more likely to apply to university and 16 percentage points more likely to attend university than their peers who do not study either of these subjects. After adjusting for background characteristics, these differences reduce to 3 percentage points in the case of applying to university and 2 percentage point for attending university. Only the first of these differences is statistically significant at the conventional 5% level; however, neither of the differences are robust to the use of the matching approach instead. Likewise, an 8 percentage point raw difference in the probability of attending a Russell Group university is reduced to a significant 2 percentage point difference after regression adjustment and further reduced to statistical insignificance when we use the matching approach.

Finally, we consider the case of whether individuals study any applied subjects. Unlike the other combinations we have considered, individuals who study applied subjects are less likely to achieve our outcome variables. Before any adjustment, we start out with a difference of 19 percentage points in the probability of applying to university, 20 percentage points in the probability of attending university, and 10 percentage points in the probability of attending a Russell Group institution. Controlling for the differences in composition of the group who study any applied subjects using a regression we find that, while the difference is much reduced to four percentage points (two percentage points in the case of Russell Group attendance), they are not eliminated. In the case of university application and university attendance, they also remain statistically significant when we restrict attention to the matched sample. This is not the case for Russell Group attendance but this could be due to few comparable individuals at this margin.

6.3.1 Continuous measure of subject choice

The main aim of this paper has been to consider what evidence there is that taking specific combinations of subjects makes a difference to later educational progression, comparing those on the margin between taking this combinations and not taking them. In general, we have found that the differences are small or not statistically significant once we account flexibly for observable differences in the individuals that take these subjects. However, perhaps it is simply the case that subject choices at age 14 simply do not affect progression to university.

To explore whether this is the case, we consider a continuous measure of academic subject selectivity (Henderson et al., 2016) and whether changes along this spectrum make a difference to the probability of progression to higher education. This measure is based on the prior academic perfor-

Table 24: Estimates of differences in university progression outcomes by continuous measure of academic selectivity of subjects studied

Outcome	**Unmatched**	**Regression**
Apply to University	0.11	0.02
	(15.78)	(2.48)
Attend University	0.11	0.01
	(14.41)	(1.08)
Attend Russell Group	0.05	0.01
	(8.70)	(1.11)
N	6000	6000

Notes: Weighted using Next Steps-provided design and attrition weights. Figures in parentheses report t-statistics from test of null hypothesis of no difference, based on standard errors calculated taking into account school-level clustering.

mance of the pupils that choose to study each subject. We assign each subject the average score in Key Stage 3 (KS3) compulsory tests at age 14 of those pupils that report they are studying that subject. KS3 tests are taken roughly contemporaneously with subject choice decisions, so they seem the most appropriate measure to use in this way. Further details, including a ranking of subjects based on this measure, are discussed by Henderson et al. (2016).

We explored a technique that extends propensity score matching to continuous 'treatments' of the kind we are considering (Bia and Mattei, 2008; Hirano and Imbens, 2005). However, it did not appear to be possible to apply this approach in this setting, largely because it relies on strong functional form assumptions. If anything, this approach appeared to exaggerate the differences. Instead, we employ the same flexible regression adjustment approach outlined in Section 6.1.1 (substituting the binary treatment indicator Treat_i for our continuous measure of subject selectivity) to estimate an average difference in outcomes across the distribution of the subject choice measure.

We report the estimated change in probability of university entry (or Russell Group attendance) for a one standard deviation change in the subject choice academic selectivity score in Table 24.

Individuals with a one standard deviation more academically selective subject mix are 11 percentage points more likely to apply to, or to attend, university. They are also five percentage points more likely to attend a Russell Group university. However, in line with the results of Henderson et al. (2016), there are big differences in the subject choice mix that young people study depending on their background characteristics. After adjusting for these differences, the difference in probability of applying to university is reduced to two percentage points, while the difference in probability of attending university, or attending a Russell Group institution, is reduced to one percentage point. Only the difference in probability of applying to university remains statistically significant.

These results provide additional context to our main findings. The association between a change in our continuous measure of academic selectivity of all subjects studied and university progression measures is small and (except in the case of application) statistically insignificant. This makes the differences in probability of going to university depending upon studying combinations of subjects that do persist even when using a matching approach (i.e. positive difference when studying a full set of subjects that would make an individual eligible for the EBacc; negative difference when studying any applied GCSEs) stand out more. It suggests that there may be a particular importance of the combinations, above and beyond that they are just more academically selective subjects.

7 Subject choices at age 14 and socioeconomic inequality in access to university

Over the past twenty years governmental efforts to promote social mobility have included widening access to higher education as a major focus. This is in an attempt to give more individuals the opportunity to benefit from the economic returns to a university degree (Walker and Zhu, 2011). Despite this, there remains a significant level of socioeconomic inequality in access to universities. Previous analyses using multiple sources of data have established that much of this gap in enrolment is explained by prior academic attainment (Chowdry et al., 2013) and by differences in application behaviour (Anders, 2012a) but that there remain some differences, particularly in access to highly competitive universities (Boliver, 2013).

However, as highlighted by through the research in this report, as well as others' research, there are important and complex patterns in the subjects individuals study during this age range. Three particularly important characteristics in explaining subject choices at this age are gender (Bell, 2001; Francis, 2000; Jin et al., 2011; Sullivan et al., 2010), prior attainment (Davies et al., 2008; Jin et al., 2011) and socioeconomic background (Davies et al., 2008; Jin et al., 2011). Previous work has considered the proximal influence of subject choice post-16. For example, in their exploration of racial inequality in university entry, Noden et al. (2014) note that differences in subject of study post-16 appear to affect university entry. Since subjects available to individuals at age 16 often depend on those that have have studied before this point, it is of interest to explore whether there are consequences of subject choices at age 14 that flow through to these same later outcomes. In particular, in this report we explore whether these overlapping socioeconomic inequalities in young people's subject choices and their university attendance interact in our research question To what extent do socioeconomic differences in the subjects individuals study from ages 14 to 16 account for residual inequality in university attendance and attendance at highly-competitive institutions (i.e. after taking into account prior attainment)?

This section explores whether differences in the subjects that individuals from different backgrounds study between ages 14 and 16 explain the remaining gap, at least in part. This complements the previous sections, which considered the direct influence of studying particular subjects on the probability of attending university.

Our results replicate previous findings of socioeconomic inequality in entry to university after conditioning on prior attainment at ages 11 and 16. Adding in controls for the subjects studied from ages 14 to 16 explains a small but significant proportion of this remaining gap. Similar results emerge when we explore inequality in entry to highly competitive universities.

7.1 Method

We use linear probability regression models (calculating heteroskedasticity- and school cluster-robust standard errors) to estimate the relationship between socioeconomic status, subjects studied, and university attendance (and, separately, attendance of a highly competitive university, specifically a member of the research intensive Russell Group). While the problems of linear probability models are well-known, these are used rather than probit or logistic binary choice regression models in order to ensure comparability across the models (Mood, 2010). Nevertheless, a similar narrative emerges if we use binary choice models instead.

Our first model (M0) includes only dummy variables indicating the quintile group of SES described in Section 3. This replicates the descriptive findings from that section, providing a 'raw' SES gap in university attendance. Furthermore, these results provide a baseline against which to judge the reduction in inequality we see when additional characteristics are added to the model. From this starting point, we estimate sequential models that highlight the overall level of socioeconomic inequality in university attendance and how much of this is explained by inequality at various points through the education system. These covariates may be thought of as 'transmission mechanisms' between SES and university attendance: to the extent that they are socially graded, their inclusion will reduce the 'raw' SES gap and provide insight on the routes through which those from more advantaged backgrounds end up being more likely to attend university.

Our first model of substantive interest (M1) adds a selection of individual- and school-level demographic characteristics that may be relevant to the relationship. These include categories of ethnicity, month of birth, government office region, number of siblings, number of older siblings and school type variables. The second and third models add individuals' prior attainment (captured using individuals scores in national tests) at age 11 (M2) and age 16 (M3). As previously noted, previous studies have found that scores in these tests explain a large proportion of the SES difference (Chowdry et al., 2013; Anders, 2012a). Nevertheless, we expect a gap to remain.

Our fourth and final model (M4) attempts to capture the importance of subjects studied between ages 14 and 16. To this end, the model includes binary indicators of each of the subjects measured in the dataset along with an additional binary indicator of whether an individual studies a full set of English Baccalaureate subjects, which the previous section indicates may have a particular association with university attendance. This EBacc indicator may be thought of as a kind of interaction, capturing the additional change in probability of attending university from having the full set of subjects beyond each individual one.

It is important to discuss the relationship between these subject measures and other variables in the model. In particular, these are the subjects in which we are capturing performance at age 16 in M3. As such, if the difficulty of the subjects also varies, subject of study could, therefore, affect performance at age 16. This could understate the importance of, for example, subjects that are both helpful for university access and more difficult to achieve high scores in. However, this issue should not affect the interpretation of socioeconomic differences in university attendance, which are the focus of this analysis, once both of the correlations have been controlled for.

7.2 Results

The top panel of Table 25 demonstrates that there are large differences in university application and attendance by the subjects that young people have studied. While 60% of the sample apply to university (attend university), 78% (67%) of those that studied EBacc subjects, 67% (55%) of those that studied two or more sciences, 66% (53%) of those who studied History or Geography, and 69% (58%) of those that studied any languages did so. Only 49% (36%) of those that studied any applied subjects did so. Similarly, while 11% of the sample go on to attend a Russell Group university, almost twice as many who studied the EBacc do so; by contrast, nearly half as many who studied any applied subjects do so. A somewhat higher proportion go on to attend a Russell Group university if they studied two or more sciences (14%), studied a foreign language (16%), or studied History or Geography (14%).

Table 25: Differences in university access by subjects studied ages 14-16

	EBacc-elig.		**2+ Sciences**		**Foreign Lang.**		**Hist./Geog.**		**Applied**		
	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	Overall
Outcomes											
Apply to University	0.51	0.78	0.43	0.67	0.45	0.69	0.49	0.66	0.68	0.49	0.60
Attend University	0.38	0.67	0.30	0.55	0.32	0.58	0.37	0.53	0.56	0.36	0.48
Attend Russell Group	0.07	0.21	0.05	0.14	0.05	0.16	0.06	0.14	0.15	0.06	0.11
Background Characteristics											
KS3 English Z-Score	-0.26	0.51	-0.44	0.19	-0.47	0.30	-0.31	0.17	0.21	-0.29	-0.00
KS3 Maths Z-Score	-0.26	0.52	-0.49	0.21	-0.44	0.28	-0.31	0.17	0.21	-0.29	-0.00
KS3 Science Z-Score	-0.27	0.54	-0.51	0.22	-0.44	0.28	-0.33	0.18	0.22	-0.30	-0.00
H/h income (£)	16,139	20,911	15,167	18,849	15,005	19,500	15,828	18,798	19,199	15,718	17,737
Male	0.49	0.48	0.48	0.50	0.55	0.45	0.45	0.51	0.50	0.48	0.49
Lone parent family	0.03	0.02	0.03	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02
Mother has degree	0.09	0.22	0.07	0.16	0.08	0.17	0.08	0.16	0.17	0.08	0.13
Father has degree	0.11	0.26	0.09	0.19	0.09	0.21	0.10	0.20	0.21	0.10	0.16
Selective School	0.02	0.14	0.02	0.08	0.00	0.10	0.03	0.08	0.09	0.02	0.06
School w/ 6th Form	0.60	0.69	0.60	0.64	0.59	0.66	0.59	0.65	0.63	0.62	0.63
Proportion	0.67	0.33	0.30	0.70	0.39	0.61	0.36	0.64	0.58	0.42	1.00
N	4,023	1,977	1,756	4,244	2,458	3,542	2,216	3,784	3,444	2,555	6,000

Notes: Column titles report combinations of subjects studied: EBacc-elig. = studied the full set of subjects required to be eligible for the English Baccalaureate performance measure; 2+ Sciences = studied at least two separate sciences or 'double award' science; Foreign Lang. = studied at least one of French, German, Italian and Spanish; Hist./Geog. = studied History, Geography or both; Applied = studied for at least one applied GCSE. Outcomes panel reports proportions achieving outcome measure by combinations of subjects studied. Background Characteristics panel reports mean (or proportion) of these characteristics by combinations of subjects studied. Weighted using Next Steps-provided design and attrition weights.

However, there is no indication that such differences can be interpreted as in any way causal. There are many differences in the characteristics of individuals who study these subject combinations, as can be seen in the lower panel of the same table and was explored in more detail by (Henderson et al., 2016). In general, we can see that individuals who study the full set of subjects required to be eligible for EBacc are, on average, from households with higher incomes than those who do not, scored higher in tests at age 14, are more likely to be in a selective school, and have parents who progressed to higher levels of education. All of these are plausibly important for explaining young people's increased probability of applying to and attending university and attending a Russell Group university (Anders, 2012a). The same broad pattern is evident for studying two or more sciences, for studying History or Geography, and for studying foreign languages, while the opposite is the case among those studying for any applied GCSEs.

Table 26: Differences in subjects studied and proportions applying to and attending university by socioeconomic status

Subject	Q1	Q2	Q3	Q4	Q5	Q1-Q5 Diff.
Academic Selectivity Score	-0.34	-0.13	-0.03	0.18	0.34	0.68
EBacc-eligible subjects	0.13	0.20	0.24	0.37	0.51	0.38
Applied subjects	0.60	0.53	0.48	0.40	0.27	-0.33
Outcome	Q1	Q2	Q3	Q4	Q5	Q1-Q5 Diff.
Attend University	0.19	0.24	0.32	0.47	0.68	0.49
Attend Russell Group	0.02	0.03	0.05	0.11	0.24	0.22

Notes: Adjusted using Next Steps-provided Wave 7 survey design, attrition and non-response weights. Sample: Wave 7 respondents with non-missing data on university attendance, constituent socioeconomic indicators, subject choice variables, and prior attainment data.

The upper panel of Table 26 reports differences in subjects studied by socio-economic status as captured through the measures discussed above. Individuals in the highest quintile group of socio-economic status are studying a mix of subjects over half a standard deviation higher more academically selective than their peers in the least advantaged quintile group. More than half of this most advantaged group study the full set of subjects required to be eligible for the English Baccalaureate (EBacc), compared to only 13% of the least advantaged fifth. By contrast, only just over a quarter of the most advantaged group study at least one applied subject, compared to 60% of the least advantaged group.

The lower panel of the same table demonstrates that there are large differences in university application and attendance by socioeconomic status. Under a fifth of the least advantaged group go on to attend university, compared with just under 70% of the most advantaged fifth. Similarly, while only 2% of the least advantaged fifth obtain a place at a Russell Group institution, almost a quarter of those in the most advantaged fifth do so. These findings are in line with previous work on inequality in access to university, whether by household income (Anders, 2012a) or other aspects of socioeconomic status (Boliver, 2013).

We report results separately for university attendance (Table 27) and Russell Group attendance (Table 28). Each reports predicted probabilities of attendance by an individuals quintile group of SES, along with the difference between the predicted probabilities between the top and bottom quintile groups. Furthermore, for each model after the first we report the reduction in this difference relative to the previous model and the statistical significance of the decrease in the gap, relative to a null hypothesis of no change.

We also report the adjusted R^2 of the model which provides us information on the additional proportion of variance in university attendance explained by the introduction of the additional characteristics. R^2 is meaningful in linear probability models and may be interpreted as the difference between the average predicted probability between the groups for whom the outcome is and is not realised (Groneau, 1998). We use adjusted, rather than unadjusted, R^2 to take into account the large number of additional covariates added to the model when accounting for subject choices.

Table 27: Regression model of university attendance

	M0	M1	M2	M3	M4
SES predicted proportions					
Q1	0.19	0.19	0.28	0.35	0.35
Q2	0.24	0.26	0.28	0.30	0.30
Q3	0.32	0.33	0.33	0.32	0.33
Q4	0.47	0.47	0.44	0.41	0.41
Q5	0.68	0.65	0.57	0.52	0.50
Q5-Q1	0.49	0.46	0.29	0.17	0.15
P > |F|	0.00	0.00	0.00	0.00	0.00
Reduction in Q5-Q1	-	0.03	0.18	0.12	0.02
P > |F|	-	0.04	0.00	0.00	0.00
Other variables in model					
Demographics		√	√	√	√
Age 11 attainment			√	√	√
Age 16 attainment				√	√
Subjects studied 14-16					√
N	7491	7488	7488	7488	7488
Adj. R^2	0.13	0.21	0.30	0.37	0.39

Notes: Weighted using Next Steps-provided design and attrition weights. Inference testing conducted using heteroskedasticity- and school cluster-robust standard errors (reported in parentheses, where relevant). Q5-Q1 reports the difference between lowest and highest quintile groups; P>|F| reports statistical significance of a joint test of all quintile groups. Reduction in Q5-Q1 reports this relative to the model to the left; P>|F| reports statistical significance of this reduction.

We begin by discussing the results for predicting university attendance in Table 27. The results for M0, in the first column, replicate those reported in 26. There is a 49%pt. difference in university attendance between the most and least advantaged fifths of the sample. Our first model including individuals' demographic characteristics and a selection of school-level covariates explains a significant proportion of variation in university attendance but makes little difference to the gap between the least and most advantaged groups: the reduction is only three percentage points.

Next, we begin to include prior attainment in our models, which we expect to explain a much larger proportion of the gap in university attendance associated with SES. Beginning with age 11 attainment (M2), we find that the difference between the most and least advantaged quintile groups is reduced to 29%pts., a reduction of 18%pts. This is a statistically significant reduction in the socioeconomic status once prior attainment to this point has been taken into account. Next, adding attainment at age 16 (M3), we see a further significant reduction in the SES gap once this prior performance has been accounted for. The difference in probability of attending university between the top and bottom quintile groups is, this time, reduced to 17%pts.. Thus far, this has replicated previous work, finding that socioeconomic inequality in access to university is, in significant part, explained by young people's prior academic attainment but that there remain significant differences, even once these have been taken into account.

Finally, we consider whether the picture changes when the subjects that individuals study are taken into account (M4). There is only a small increase in adjusted R^2 when these indicators are added to the model. Nevertheless, there is a small but statistically significant narrowing of inequality (two percentage points), driven by a reduction in the conditional probability of attending university among the most advantaged quintile group. This suggests that, at most, changes in the subjects that individuals study between these points in time can make, at most, only a small difference to narrowing socioeconomic differences in university going.

Table 28: Regression model of Russell Group attendance

	M0	M1	M2	M3	M4
SES predicted proportions					
Q1	0.02	0.03	0.06	0.08	0.08
Q2	0.03	0.04	0.05	0.05	0.06
Q3	0.05	0.05	0.05	0.05	0.05
Q4	0.11	0.10	0.09	0.08	0.08
Q5	0.24	0.22	0.19	0.18	0.17
Q5-Q1	0.22	0.19	0.13	0.10	0.09
P > |F|	0.00	0.00	0.00	0.00	0.00
Reduction in Q5-Q1	-	0.04	0.06	0.03	0.01
P > |F|	-	0.00	0.00	0.00	0.00
Other variables in model					
Demographics		$\surd$	$\surd$	$\surd$	$\surd$
Age 11 attainment			$\surd$	$\surd$	$\surd$
Age 16 attainment				$\surd$	$\surd$
Subjects studied 14-16					$\surd$
N	7481	7478	7478	7478	7478
Adj. R^2	0.08	0.13	0.16	0.18	0.19

Notes: Weighted using Next Steps-provided design and attrition weights. Inference testing conducted using heteroskedasticity- and school cluster-robust standard errors (reported in parentheses, where relevant). Q5-Q1 reports the difference between lowest and highest quintile groups; P>|F| reports statistical significance of a joint test of all quintile groups. Reduction in Q5-Q1 reports this relative to the model to the left; P>|F| reports statistical significance of this reduction.

Next, we repeat the analysis this time considering probability of attending a highly-competitive Russell Group university. Again, M0 replicates the descriptive differences reported in Table 26. Overall, the levels are much lower, however there are still big differences, with those in the most advantaged fifth 12 times (22 %pts.) more likely to attend one of these high-status universities than their peers in the bottom SES quintile group.

Including demographic characteristics (M1) makes a slightly larger relative difference in the case of inequality in access to a high-status university, suggesting these ethnicity and school type measures may partly explain the raw SES differences we report. However, as with university access more generally, a big proportion inequality in access to high-status universities is explained by performance in tests at age 11 (M2). The gap between the most and least advantaged quintile groups narrows by 6%pts. (just under a third of the remaining gap) when this is taken into account. This narrows by a further three percentage points once performance at age 16 are taken into account (M3). So far, all of these reductions have been statistically significant but have left significant remaining inequality.

Finally, we consider the difference made by subjects studied (M4). Again, the broad picture is of a small but statistically significant reduction in inequality in attending a highly competitive university. As with inequality in attending any university, this leaves a significant proportion of inequality explained by the factors included in the model.

8 Conclusions

This section describes findings from each of the four main strands of the report and draws out overall conclusions relevant to policymakers and practitioners.

8.1 Individual predictors

There are some common predictive factors across the outcomes we have looked at. As expected, prior attainment is consistently positively associated with taking selective subjects, facilitating subjects, STEM subjects and EBacc-eligble subjects, while it is negatively associated with taking applied GCSEs. This suggests that high achieving students are directed towards particular subjects. However, since previous research has highlighted the socioeconomic stratification of attainment of pupils by age 14, this pattern is also consistent with the existence of primary effects of family background, i.e. acting through prior attainment, for all of our metrics for subject choice at age 14 in England. However, there is also evidence of important secondary effects, since parental socio-economic background remains an important predictor even once prior attainment is held constant. The secondary effects are greatest in the case of subject selectivity, and weakest in the case of STEM. The presence of strong secondary effects of stratification for subject selectivity, facilitating subjects, vocational subjects, and EBacc confirms the potential for curriculum choice at 14 to exacerbate inequalities rather than simply reflect existing inequalities. The pattern of ethnic differentials across our outcomes is not completely consistent, suggesting that ethnic patterns are rather sensitive to the particular curriculum categorisation used. For example, ethnic minority pupils are, broadly speaking, advantaged in terms of facilitating subjects, but there is no such clear pattern for subject selectivity or for Ebacc-eligible subjects. Girls have lower odds of taking three or more STEM subjects and higher odds of taking Applied GCSEs, however we found no significant gender difference in EBacc or facilitating subjects.

The results point to an important school effect which requires more research to assess the roles of the subjects offered within schools and informal school policies which may influence which students are allowed to take particular subjects. The present analysis finds that grammar school status is positively associated with doing EBacc-eligble subjects, STEM and facilitating subjects (although for STEM and facilitating subjects the results are not significant) and negatively associated with doing Applied GCSEs; this is not entirely explained by the higher prior attainment of those who attend such schools. Our analytical sample excludes private schools, but given the difference between comprehensive and grammar schools we can speculate that subject choice would also vary according to private school status, particularly as we find that attending a single-sex school is positively associated with doing a more selective curriculum and EBacc-eligible subjects. We find that the proportion of free school meals-eligible students in the school is negatively associated with all subject choice metrics except in the case of applied GCSEs, which is not statistically significant. This finding accords with the wider literature showing that school SES matters for individual pupil outcomes (Marks, 2015; Perry and McConney, 2010; Caldas and Bankston, 2012).

8.2 Schools

The analysis of the extent to which individuals' decisions are associated with the school they attend also sheds light on the importance of constraints on schools, especially how schools' provision (and, hence, the options open to their pupils) is shaped by their composition in terms of academic attainment, socioeconomic background and gender mix.

The results replicate previous findings that young people's prior attainment, socioeconomic background, and gender are all associated with the subjects they study at age 14-16. However, our novel contribution is to consider the separate associations between the composition of their school in these terms and subjects studied. We find that individuals in schools with more advantaged intakes are more likely to study more academically selective subjects, even after conditioning on individuals' own SES. Individuals' prior attainment is associated with studying more academically selective subjects as, again, is the prior attainment of the school more generally. Male students are less likely to study academically selective subjects than their female counterparts, although there is also evidence of higher subject selectivity in single-sex schools of either type.

Overall, schools explain about a third of the variation in the academic selectivity of the subjects that young people study; once we take into account the demographics of the school this is reduced to closer to a quarter. Constraints that schools may face on the subjects that they can viably offer are also highlighted by the negative association between being in a local authority in which a significant proportion of pupils attend selective schools and the academic selectivity of subjects that individuals study.

There are some important differences when we consider some examples of whether individuals study for specific sets of subjects that it has been argued may be important for future academic outcomes. The odds of studying any applied subjects is lower among pupils in schools with a higher prior attaining intake. While it is the case that in most single sex schools pupils are more likely to study a more academically selective set of subjects, for studying triple sciences this is only true in all-male schools and not all-female ones.

Overall, this paper has highlighted the important role that schools seem to play in many subject choice decisions, with significant variation in subjects studied attributable to the school-level. However, it also highlights that, in many cases, what we might see as schools' actions are, in fact, strongly affected by constraints on what they can viably offer, for example as due to their intake in terms of prior attainment, gender and socioeconomic status. The socio-economic composition of the school is a powerful predictor of individual choices in our models, even controlling for the academic composition of the school. In fact, the SES composition of the school has a similar level of effect on subject choices as the individuals family SES.

We suggest two potential mechanisms. First, schools may try to offer a curriculum which they deem appropriate for the socio-economic composition of the school; in other words, schools may deliberately take school SES into account when choosing the subjects offered. Second, we know that schools serving disadvantaged children face difficulties in recruiting and retaining highly qualified staff (Lupton, 2005; Lupton and Thrupp, 2012; Ofsted, 2013) particularly in shortage areas such as languages and sciences. This may constrain the curriculum that schools with disadvantaged intakes are able to offer. The strength of the school SES effect is surprising, and has potential policy implications, as it suggests an additional pathway through which school SES influences academic outcomes for pupils.

8.3 Combinations

Using rich survey data collected about a representative cohort of young people from England, we estimated the effect on university entry of studying specific sets of subjects that have been of particular particular interest to policymakers in recent years. We did so using both regression modelling and propensity score matching to test the robustness of the results to each of these approaches. The aim of these methods is to produce estimates that specifically compare very similar individuals who are on the margin between studying a set of subjects or not.

While there are large raw differences in the probability of university attendance by subjects studied, once differences in the characteristics of individuals who study such subjects are taken into account, the remaining differences are small or non-existent. There is some evidence of a positive effect (3-5 percentage points) of studying the full suite of English Baccalaureate subjects and a negative effect of a similar magnitude of studying any applied GCSE subjects. We also produced estimates of the change in probability of applying to or attending university, or attending a Russell Group university, associated with a general increase in the academic selectivity of the subjects that an individual studies. The differences are also small and, except in the case of applying to university, not statistically significant.

With respect to the significant differences, we should keep in mind that to regard these results as truly causal we need to be satisfied that there are no unobserved differences (i.e. driven by factors that we could not include in the propensity score model) between individuals who did study such subjects and those who did not that could be driving the results. Given the relatively small differences we find after taking the observed differences into account, it would not require large unobserved differences for these results to be overturned. It is striking that, while differences associated with a continuous change in the academic selectivity of subjects studied become insignificant when controlling for background characteristics, differences in the probability of university attendance by whether or not young people studied the full set of EBacc subjects and by whether or not they studied any applied subjects remain significant when controlling for the same set of background characteristics. This suggests that there may be a particular importance of the combinations, above and beyond that they are just more academically selective subjects.

8.4 Inequality

We have also explored the relationship between socioeconomic status, subjects studied from ages 14-16, and university attendance (including attendance of highly competitive universities). This brings together previous research on both the socioeconomic correlates of subject choice (Henderson et al., 2016) and inequality in access to university Anders (2012a); Boliver (2013); Chowdry et al. (2013) in England.

We provide new evidence on whether subjects of study during this period help to explain the remaining socioeconomic gap in university attendance once attainment at age 16 has been accounted for. We find evidence that they do explain a small, but significant, part of the gap but that much more remains unaccounted for. This finding is also true for the inequality in attending highly competitive universities.

8.5 Overall implications

Overall, the findings of this project highlight a number of important implications for schools and policymakers. They highlight that we should be sceptical of considering young people's subjects of study purely in terms of 'choice' (Woods, 1976). They are, at most, constrained choices, potentially both for individuals and for schools.

We found that even after controlling for prior attainment and school difference, young people from advantaged households take more selective subjects, have higher odds of doing three or more facilitating subjects and EBacc-eligible subjects and lower odds of taking Applied GCSEs than less advantaged young people. This is likely to be partly a result of direct forms of support from parents with higher socio-economic status which leads to difference in subject choice, but we also found evidence to support an indirect effect via school differences. We found that there were important differences by school characteristics, which may be a result of differential opportunities, subjects offered and within school policies.

The findings regarding subjects studied in schools within selective local authorities suggest that expanding selective education will both increase inequality in, and decrease the average level of, academic selectivity of subjects that young people study.

We find that the seemingly large differences in university progression associated with the subjects young people study from ages 14 to 16 often seem to be small, at most, once we take into account differences in the kinds of people who study these subjects. Why might this be the case? Beyond the removal of the influence of background characteristics, it could be that differences in subjects studied at this stage are simply swamped by decisions young people make in the following two years.

The results for studying the full set of EBacc subjects and for studying any applied subjects do show residual associations with university attendance, suggesting the view that they may have particular importance is not without merit, a finding that concords with other research focussing on subject choice at a later point in individuals' educational careers (Dilnot, 2016). Nevertheless, it is important to emphasise that the differences are still not large, suggesting that the weight that has been placed on the EBacc by policymakers has been exaggerated.

Our findings suggest that if young people from different socioeconomic backgrounds were studying a more similar curriculum between ages 14 and 16 it would be unlikely to make much of difference to the inequality in university entry highlighted by previous studies. This does not mean that ensuring pupils have the same opportunities to choose their curriculum post-14 regardless of their background is not important. We may regard this as important in itself for reducing socioeconomic differences in earlier educational trajectories, and also having the potential to make a difference to inequality in university going at the margin. However, we certainly should not regard reforms in this space as any kind of 'silver bullet'.

Bibliography

Alcott, B. (2017). Does Teacher Encouragement Influence Students' Educational Progress? A Propensity-Score Matching Analysis. Research in Higher Education. doi:10.1007/s11162-017-9446-2.

Anders, J. (2012a). The Link between Household Income, University Applications and University Attendance. Fiscal Studies, 33(2):185–210. doi:10.1111/j.1475-5890.2012.00158.x.

Anders, J. (2012b). Using the Longitudinal Study of Young People in England (LSYPE)/"Next Steps" for research into Higher Education access. DoQSS Working Paper 12-13, Institute of Education, University of London.

Bell, J. F. (2001). Patterns of Subject Uptake and Examination Entry 1984-1997. Educational Studies, 27(2):201–219. doi:10.1080/03055690120050428.

Bia, M. and Mattei, A. (2008). A Stata package for the estimation of the dose-response function through adjustment for the generalized propensity score. The Stata Journal, 8(3):354–373.

Boliver, V. (2013). How fair is access to more prestigious UK universities? British Journal of Sociology, 64(2):344–364. doi:10.1111/1468-4446.12021.

Borghans, L., Golsteyn, B. H. H., and Stenberg, A. (2013). Does Expert Advice Improve Educational Choice? IZA Discussion Paper IZA DO No. 7649, Institute for the Study of Labor.

Breen, R. and Yaish, M. (2006). Testing the Breen-Goldthorpe Model of Educational Decision-Making. In Morgan, S. L., Grusky, D. B., and Fields, G. S., editors, Mobility and Inequality: Frontiers of Research from Sociology and Economics, pages 232–258. Stanford University Press.

Chevalier, A. (2004). Parental education and child's education: a natural experiment. CEE Discussion Paper CEE DP 40, Centre for the Economics of Education, London School of Economics and Political Science.

Chmielewski, A. K. (2014). An International Comparison of Achievement Inequality in Within- and Between-School Tracking Systems. American Journal of Education, 120(3):293–324. doi:10.1086/675529.

Chowdry, H., Crawford, C., Dearden, L., Goodman, A., and Vignoles, A. (2013). Widening participation in higher education: analysis using linked administrative data. Journal of the Royal Statistical Society: Series A, 176(2):431–457. doi:10.1111/j.1467-985X.2012.01043.x.

Codiroli, N. (2015). Inequalities in students' choice of STEM subjects. CLS Working Paper 2015/6, Centre for Longitudinal Studies, UCL Institute of Education.

Coleman, J. S., Campbell, E. Q., McPartland, J., Mood, A. M., Weinfeld, F. D., and York, R. L. (1966). Equality of Educational Opportunity. National Center for Educational Statistics, U.S. Department of Health, Education, and Welfare, Washington, DC.

Davies, P., Telhaj, S., Hutton, D., Adnett, N., and Coe, R. (2008). Socioeconomic background, gender and subject choice in secondary schooling. Educational Research, 50(3):235–248. doi:10.1080/00131880802309358.

Dilnot, C. (2016). How does the choice of A-level subjects vary with students' socio-economic status in English state schools? British Educational Research Journal. doi:10.1002/berj.3250.

Duquet, N., Glorieux, I., Laurijssen, I., and Van Dorsselaer, Y. (2010). An unfinished job? The effect of subject choice and family formation processes on labour market outcomes of young men and women. Journal of Education and Work, 23(4):319–338. doi:10.1080/13639080.2010.484800.

Ermisch, J. and Pronzato, C. (2010). Causal Effects of Parents' Education on Children's Education. ISER Working Paper 2010-16, Institute of Social and Economic Research, University of Essex.

Francis, B. (2000). The Gendered Subject: Students' subject preferences and discussions of gender and subject ability. Oxford Review of Education, 26(1):35–48. doi:10.1080/030549800103845.

Gibb, N. (2011). Speech by Nick Gibb to the Specialist Schools and Academies Trust.

Goldthorpe, J. H. and McKnight, A. (2004). The Economic Basis of Social Class. CASEpaper 80, Centre for Analysis of Social Exclusion, London School of Economics.

Groneau, R. (1998). A useful interpretation of R2 in binary choice models (or, have we dismissed the good old R2 prematurely). Working paper, Industrial Relations Section, Princeton University.

Hauser, R. M. (1994). Measuring Socioeconomic Status in Studies of Child Development. Child Development, 65(6):1541–1545. doi:10.2307/1131279.

Havari, E. and Savegnago, M. (2014). The causal effect of parents' schooling on children's schooling in Europe. A new IV approach. CEIS Tor Vergata Research Paper Vol. 12, Issue 5, No. 315, Cenre for Economic and International Studies Tor Vergata.

Henderson, M., Sullivan, A., and Anders, J. (2016). Social Class, Gender and Ethnic Differences in Subjects Taken at Age 14. CLS Working Paper, Centre for Longitudinal Studies, UCL Institute of Education.

Henderson, V., Mieszkowski, P., and Sauvageau, Y. (1978). Peer group effects and educational production functions. Journal of Public Economics, 10(1):97–106. doi:10.1016/0047-2727(78)90007-5.

Hirano, K. and Imbens, G. M. (2005). The propensity score with continuous treatments. In Gelman, A. and Meng, X.-L., editors, Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family, chapter 7, pages 73–84. John Wiley and Sons, Ltd., Chichester, UK.

Hodgson, A. and Spours, K. (2008). 14-19 Education and Training: Curriculum, Qualifications and Organisation. Sage.

Hoxby, C. (2000). Peer effects in the classroom: learning from gender and race variation. NBER Working Paper 7867, National Bureau of Economic Research.

Hoxby, C. and Turner, S. (2013). Expanding College Opportunities for High-Achieving, Low Income Students. SIEPR Discussion Paper 12-014, Stanford Institute for Economic Policy Research, Stanford University.

Iannelli, C. (2013). The role of the school curriculum in social mobility. British Journal of Sociology of Education, 34(5-6):907–928. doi:10.1080/01425692.2013.816031.

Imbens, G. M. and Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Intro Cambridge University Press, New York, NY.

Jenkins, S. P. and Schluter, C. (2002). The effect of family income during childhood on later-life attainment: evidence from Germany. IZA Discussion Paper 604, Institute for the Study of Labor.

Jin, W., Muriel, A., and Sibieta, L. (2011). Subject and course choices at ages 14 and 16 amongst young people in England: insights from behavioural economics. DfE Research Report DFE-RR160, Department for Education.

Kolenikov, S. and Angeles, G. (2009). Socioeconomic Status Measurement with Discrete Proxy Variables: Is Principal Component Analysis a Reliable Answer? Review of Income and Wealth, 55(1):128–165. doi:10.1111/j.1475-4991.2008.00309.x.

Lavy, V. and Schlosser, A. (2011). Mechanisms and Impacts of Gender Peer Effects at School. American Economic Journal: Applied Economics, 3(2):1–33. doi:10.1257/app.3.2.1.

Lupton, R. (2005). Social justice and school improvement: improving the quality of schooling in the poorest neighbourhoods. British Educational Research Journal, 31(5):589–604. doi:10.1080/01411920500240759.

Lupton, R. and Thrupp, M. (2012). Headteachers readings of and responses to disadvantaged contexts: evidence from English primary schools. British Educational Research Journal, pages 1–20. doi:10.1080/01411926.2012.683771.

Marks, G. N. (2006). Are between- and within-school differences in student performance largely due to socio-economic background? Evidence from 30 countries. Educational Research, 48(1):21–40. doi:10.1080/00131880500498396.

Marks, G. N. (2015). Are school-SES effects statistical artefacts? Evidence from longitudinal population data. Oxford Review of Education, 41(1):122–144. doi:10.1080/03054985.2015.1006613.

McCrone, T., Morris, M., and Walker, M. (2005). Pupil Choice at Key Stage 3 - Literature Review. Nfer report, National Foundation for Educational Research.

McDool, E. (2017). Neighbourhood Effects on Educational Attainment: Does Family Background Influence the Relationship? SERPS 2017002, The University of Sheffield Department of Economics.

Mendolia, S. and Walker, I. (2014). The Effect of Personality Traits on Subject Choice and Performance in High School: Evidence from an English Cohort. IZA Discussion Paper 8269, Institute for the Study of Labor.

Mood, C. (2010). Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. European Sociological Review, 26(1):67–82. doi:10.1093/esr/jcp006.

Mortimore, P. (1988). School Matters. University of California Press.

Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. Econometrica, 46(1):69–85. doi:10.2307/1913646.

Noden, P., Shiner, M., and Modood, T. (2014). University offer rates for candidates from different ethnic categories. Oxford Review of Education, 40(3):349–369. doi:10.1080/03054985.2014.911724.

Ofsted (2013). The report of Her Majesty's Chief Inspector of Education, Children's Services and Skills 2013/14: Schools. Annual Report, Ofsted.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika, 44(4):443–460. doi:10.1007/BF02296207.

Open Public Services Network (2015). Lack of options: how a pupil's academic choices are affected by where they live. OPSN Report, Open Public Services Network.

Proud, S. (2014). Girl Power? An analysis of peer effects using exogenous changes in the gender make-up of the peer group. Regional and Sectoral Economic Studies, 14(3):5–18.

Rose, D. and Pevalin, D. J. (2001). The National Statistics Socio-economic Classification: Unifying Official and Sociological Approaches to the Conceptualisation and Measurement of Social Class. ISER Working Paper 2001-4, Institute for Social and Economic Research, University of Essex.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55. doi:10.1093/biomet/70.1.41.

Russell Group (2013). Informed Choices: A Russell Group guide to making decisions about post-16 education. Russell group guide.

Rutter, M. (1982). Fifteen Thousand Hours: Secondary Schools and Their Effects on Children. Harvard University Press.

Shadish, W. R. (2012). Propensity score analysis: promise, reality and irrational exuberance. Journal of Experimental Criminology, 9(2):129–144. doi:10.1007/s11292-012-9166-8.

Smith, D. J., Tomlinson, S., Bonnerjea, L., Hogarth, T., and Thomes, H. (1989). The School Effect: A Study of Multi-Racial Comprehensives. Policy Studies Institute, London.

Sullivan, A. (2009). Academic self-concept, gender and single-sex schooling. British Educational Research Journal, 35(2):259–288. doi:10.1080/01411920802042960.

Sullivan, A., Joshi, H., and Leonard, D. (2009). Single-Sex Schooling and Academic Attainment at School and Through the Lifecourse. American Educational Research Journal, 47(1):6–36. doi:10.3102/0002831209350106.

Sullivan, A., Zimdars, A., and Heath, A. (2010). The social structure of the 14–16 curriculum in England. International Studies in Sociology of Education, 20(1):5–21. doi:10.1080/09620211003655572.

Tripney, J., Newman, M., Bangpan, M., Niza, C., MacKintosh, M., and Sinclair, J. (2010). Subject Choice in STEM: Factors Influencing Young People (aged 14-19) in Education. A systematic review of the UK literature. Wellcome trust report, EPPI Centre, Social Science Research Unit, Institute of Education, University of London.

Van Houtte, M. (2004). Gender context of the school and study culture, or how the presence of girls affects the achievement of boys. Educational Studies, 30(4):409–423. doi:10.1080/0305569042000310336.

Vidal Rodeiro, C. L. (2009). Some issues on the uptake of Modern Foreign Languages at GCSE. Statistical Report Series 10, Cambridge Assessment.

Walker, I. and Zhu, Y. (2011). Differences by degree: Evidence of the net financial rates of return to undergraduate study for England and Wales. Economics of Education Review, 30(6):1177–1186. doi:10.1016/j.econedurev.2011.01.002.

Willms, J. D. (1986). Social Class Segregation and Its Relationship to Pupils' Examination Results in Scotland. American Sociological Review, 51(2):224–241.

Woods, P. (1976). The Myth of Subject Choice. British Journal of Sociology, 27(2):130–149.