

## **Matching in the dark? Inequalities in student to degree match**

Centre for Education Policy and Equalising Opportunities (CEPEO)

**Stuart Campbell, Lindsey Macmillan, Richard  
Murphy and Gill Wyness**

**Working Paper No. 20-01**

**January 2020**

## Disclaimer

Any opinions expressed here are those of the author(s) and not those of the UCL Institute of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

CEPEO Workings Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

## Highlights

- Research from the US has shown that a significant proportion of students are 'undermatched' – attending courses that are less selective than might be expected given their academic attainment. We study this – and the inverse phenomenon of 'overmatch' – in the UK for the first time.
- We create two measures of student to degree course match, measuring course quality according to the attainment of students on the course, and based on the earnings of graduates of the course.
- We find significant under- and over-match in the UK, with 30% of students mismatched when we measure course quality based on attainment, and almost half of students mismatched when we measure course quality based on graduate earnings.
- We find that students from lower socio-economic groups systematically undermatch for both measures across the entire distribution of achievement. And, while women attend courses that are just as academically selective as men, high-attaining women systematically attend courses that have lower expected earnings than men.
- Given the wage returns associated with more selective courses, these findings have important implications for equity and social mobility, and the gender pay gap. They also have implications for schools in their information, advice and guidance strategies.

## Why does this matter?

Simply encouraging low income students to attend university isn't enough to equalize opportunities. We need to pay attention to the types of courses these students are attending.

We should target low income and female students with information – such as on the courses that match their attainment profile, and the earnings associated with different courses both before, and at the application stage

# Matching in the Dark? Inequalities in student to degree match

Stuart Campbell<sup>a</sup>, Lindsey Macmillan<sup>a</sup>, Richard Murphy<sup>b,c</sup> and Gill Wyness<sup>a,b</sup>

December 2019

## Abstract:

This paper examines inequalities in the match between student quality and university quality using linked administrative data from schools, universities and tax authorities. We analyse two measures of match at the university-subject (degree) level, based on student academic achievement, and graduate earnings. We find that students from lower socio-economic groups systematically undermatch for both measures across the entire distribution of achievement, with particularly stark socio-economic gaps for the most undermatched. This is in a setting with no up-front tuition fees and a generous financial aid system. We show that distance is not a driving factor, but that secondary schools play a key role in generating these gaps. While there are negligible gender gaps in academic match, high-attaining women systematically undermatch in terms of expected earnings, largely driven by subject choice.

**JEL classification:** I22, I23, I28

**Keywords:** higher education, educational economics, college choice, mismatch, undermatch

**Acknowledgements:** We thank Paul Gregg, Sandra McNally, John Friedman, Peter Bergman and Leigh Linden for helpful comments, as well as seminar participants at Austin, Columbia, CEP, Cornell, ISER, and Queen's Belfast, workshop participants at Catanzaro, IAB, and York, and conference participants at APPAM, EALE, ESPE, RES and SOLE.

Wyness and Macmillan acknowledge Nuffield Foundation funding (172585).

Correspondence: g.wyness@ucl.ac.uk

<sup>a</sup> UCL Institute of Education

<sup>b</sup> Centre for Economic Performance, London School of Economics and Political Science

<sup>c</sup> University of Texas at Austin

## 1. Introduction

Increasing enrolments in higher education (HE) is a preoccupation of governments around the world. As a result, much academic research has been devoted to examining policies intended to increase participation by relaxing credit constraints (Carneiro & Heckman 2002, Lochner & Monge-Naranjo 2011, Murphy et al 2019), providing better information (Hoxby & Turner 2015, McGuigan et. al 2016, Dynarski et at, 2018,) or improving prior academic achievement (Avery, 2013, Chowdry et al, 2013). However, less attention has been given to the types of universities and degrees students enrol in once they decide to continue with their education.

This is an issue of critical importance given the recent evidence showing the gains from students being well-matched to their degrees, with students who over or undermatch underperforming (Arcidiacono and Lovenheim, 2016; Dillon and Smith, 2019 respectively). The existence of these complementarities between degree and student quality, mean that matching students to degrees has large potential impacts on the aggregate returns to higher education for society.<sup>1</sup> A market in which there is mismatch would imply that there are inefficiencies.<sup>2</sup>

How efficient is the matching market in the higher education sector? And are some students systematically mismatching? This paper takes a step forward in answering these questions by using administrative data from all state schools, universities and the tax authority in England. From this we track the entire cohort of 140,000 students from school to university to construct measures of student to university-subject (henceforth described as “degree”) match. Using these measures, we document the extent of mismatch and the types of students that are systematically mismatching.

We create two measures of match. For both, we rank students nationally based on their end of secondary school qualifications. We also rank degrees nationally, first, according to the qualifications of the median student on each degree, and second, according to the median earnings of previous graduates on the degree. We create our measures of match by taking the difference between the percentile ranking of the student and the degree. The advantage of our

---

<sup>1</sup> Even in the absence of penalties to overmatching, the existence of capacity constraints and undermatch penalties will deliver the same conclusion. Explicitly, if some students are overmatching, given the capacity constraints, this will result in some students undermatching.

<sup>2</sup> If, on the other hand, we assume there are no complementarities between students and courses, a well-functioning market would be one in which students attend the course with the highest return, given their characteristics (including attainment). Here, mismatch again implies the existence of inefficiencies, such as frictions in the application and enrolment process, or that students are badly informed about course quality. However, student preferences could also create the appearance of mismatch, which would be problematic since preferences cannot typically be ruled out.

approach is that it provides a transparent and continuous measure of match which can be used in many settings. Moreover, defining match to degree on the basis of earnings is a new addition to the literature, which allows us to shed light on previously undocumented large disparities in match.

We use these two measures to document socio-economic status (SES) and gender differences in match, taking three distinct approaches. First, we plot student qualification percentile against degree quality percentile for students by SES and gender. Plotting the raw data in this way imposes no functional form assumptions on the data, and presents the extent of match throughout the achievement distribution. Second, we estimate the average SES/gender match gaps conditional on individual characteristics and prior achievement across the entire distribution of achievement – an important shift in the literature, which has typically focused on high achievers. Finally, we implement unconditional quantile regressions (UQR) across the distribution of match to determine whether these mean effects are masking larger SES and gender gradients for those who are very mismatched. The combination of these approaches allows us to explore hidden non-linearities across the entire academic achievement distribution, and reveals several important findings.

We find sizeable socio-economic gaps in academic and earnings match across the achievement distribution, with low SES students consistently undermatching, attending degrees with lower attaining peers and lower expected earnings than their richer counterparts. These gaps remain after conditioning on a set of individual demographics and a complete history of prior test scores. In the top quintile of the achievement distribution, disadvantaged students are 8 percentiles lower matched than their more advantaged counterparts. This corresponds to the difference between studying economics at the London School of Economics (ranked 5<sup>th</sup> in the Times Higher UK university rankings) versus Exeter (ranked 18<sup>th</sup>). The largest inequalities are not found at the top of the achievement distribution, but around the 90<sup>th</sup> percentile. These disadvantaged students are 11 percentiles lower matched than their more advantaged counterparts. We find little evidence that these gaps are driven by the subjects that people study at university. Even when they have similar prior achievement, and are studying similar degree subjects, low SES students study at lower ranked (in terms of both achievement and earnings) institutions. This is consistent with recent research from Chetty et al. (2017) which documents the ‘missing middle’ – middle class students with high test scores who are under-represented at the most selective US colleges.

The existence of mismatch in the system implies that there are inefficiencies at play. Research to date has highlighted the role of credit constraints, geographical isolation and

information inequalities as the key drivers of mismatch, particularly among low SES students (Dillon and Smith, 2018; Hoxby and Avery, 2012). We can rule out the first two of these as explanations of mismatch in the system we study. Credit constraints play a minimal role in the UK; practically all university degrees charge the maximum tuition fees allowable – meaning there is no variation in fees across degrees, so poorer students cannot make a price-quality trade-off – and all students have access to income-contingent loans that cover the entirety of the tuition fees plus loans for living expenses<sup>3</sup>. This may explain why we observe less mismatch in the UK than in the US, where there is large variation in fees and such loans are not widely used<sup>4</sup>. We find that geography has little impact on the SES match gap. In our context, students only have to travel short distances to find a well-matched course, with the average distance to a well-matched course in England a little under 9 miles.<sup>5</sup> On average low SES students attend colleges closer to home, but conditioning on distance to either university attended or a well-matched degree does not impact the match parameters significantly. That low SES students are systematically undermatching despite the lack of geographic or financial constraints, means that there are other structural or social factors at work that the research to date has not fully explored.

We show that a major determinant of SES gaps in mismatch is high school attended. The SES match gap for students from the same school is reduced by up to 79 percent (down to 2 percentiles) with the inclusion of school fixed effects. This implies that factors correlated with high school such as peers, school resources, information inequalities including careers advice and guidance, and parental sorting play an important role in student match.

In addition to SES match, we provide facts about the previously undocumented gender gaps in student to degree match. In contrast to the large SES gaps in academic match, we find only modest differences in academic match between males and females. Meaning that males and females with a given set of qualifications are enrolling in courses with similar entry standards. However, by stark contrast, we find sizeable gender gaps in earnings match. After accounting for prior test scores and demographics, high-attaining women attend degrees around 8 percentiles lower in associated earnings than men - this gap is the equivalent of £25,800 per year for those degrees at the top of the median earnings distribution. This highlights that women are attending degrees that are as academically competitive as their male peers, but that have

---

<sup>3</sup> Combined tuition fee and maintenance loan take-up was 86% in 2010/11 (when eligible amounts were similar to the cohort we study) (Bolton, 2019)

<sup>4</sup> See Barr et al (2017) for a discussion.

<sup>5</sup> The UK has 3.47 universities ranked in the QS top 1000 ranking, per 10,000 km<sup>2</sup> compared to 0.17 in the US.

substantially lower average earnings. We find that almost the entire of the gender gap in earnings can be accounted for by degree subject choice (rather than university attended), with women more likely to attend degrees such as Creative Arts and English – which are academically selective, but have typically lower earnings.

Our paper makes several key contributions to the emerging academic literature on the match between student achievement and college quality. Existing papers on mismatch, and the HE inequality literature in general have typically focused on high-achieving low-income students, using a binary measure of undermatch (Hoxby and Avery, 2012; Black, Cortes and Lincove, 2015) or examined mismatch at different points in the distribution (Dillon and Smith, 2017). In contrast, we create continuous measures of mismatch, and present estimates across the distribution of achievement. Our focus on the entire distribution of achievement is a non-trivial shift in this literature. Simply focusing on high achievers obscures more dramatic undermatch among those in the 70-90th percentiles of the skills distribution. Our results imply that efforts to reduce mismatch should be expanded beyond just the highest attainers.

The continuous nature of our measures in conjunction with our large dataset also makes it possible to make a novel contribution to the literature by examining the nature of mismatch at its extremes through unconditional quartile regression (UQR). Standard OLS estimates of mismatch under-sell the importance of SES and gender gaps for the most under-matched students. We estimate OLS conditional SES earnings gap of high achieving students to be 8 percentiles, but for the most under-matched the SES gaps are as large as 27 percentiles. For gender, the earnings match-gap among the most qualified students is 16 percentiles (compared to 8 percentiles at the mean).

We are also the first to study mismatch on the basis of degree earnings potential. Previous studies have measured university quality based on entry qualifications of the students at that institution (Hoxby and Avery, 2012; Light and Strayer, 2000) or a composite of institution quality measures (Dillon and Smith, 2017; Smith, Pender and Howell, 2013). Measuring university quality on the basis of graduate earnings is important for understanding the role of match in intergenerational mobility. Our finding that talented low SES students are enrolling in degrees with lower returns undermines the potential for higher education to have a positive impact on social mobility.

A final contribution is that we can study mismatch at university subject (degree) level. All existing studies of mismatch have been unable to untangle the role of university subject/major as a factor in match. This, in conjunction with our new measure of earnings match, allows us to highlight a large and undocumented gender match gap. Our finding that



talented women enrol in subjects which command lower returns than equally talented men is relevant for the much documented gender pay gap. It also implies that interventions aimed at improving match should be targeted at women as well as disadvantaged students.

The remainder of this paper proceeds as follows. Section 2 describes our institutional setting, the dataset, and the methods we use to create our indices of undermatch. Section 3 presents our results from the three approaches, while Section 4 presents robustness tests. Section 5 explores potential drivers of undermatch, and Section 6 concludes.

## **2. Data and Methods**

### **2.1 Institutional setting**

We analyse inequalities of match in the UK context, which provides some perspective on the findings from the predominantly American literature. While other studies of mismatch have pointed to the role of finance as a potential driver (Hoxby and Avery, 2012), UK students face far fewer financial barriers. There are no upfront costs in the UK system - all college fees and living costs are covered by income-contingent loans which are repaid upon graduation once the graduate is earning over a certain level (Murphy, Scott-Clayton and Wyness, 2019). Moreover, there is little price variation between institutions or subjects meaning that students do not face a trade-off between quality and price which may cause them to mismatch. A final feature of the UK system is that it has a centralised applications system (the University and College Applications Service, or UCAS) which is easy to access and navigate and is used by the vast majority of university applicants. Students are provided with standardized information on all the degrees including typical grade requirements, and can apply for up to five degrees paying a single application fee of £24. Thus, the finding of substantial student to university mismatch even in a system with few financial barriers, relatively low costs, and streamlined application system is important, pointing to other possible reasons for this mismatch.

As in the US, students are still likely to face information constraints, however. The structure of the UK education system means that students make a number of crucial choices about their education path as early as age 13/14. At this age, students choose the types of qualifications and, crucially, subjects that they will study in their final two years of compulsory schooling, most often for 10 General Certificates of Secondary Education (GCSEs). Those who stay on after the compulsory schooling age face another set of important decisions regarding their qualification and subject choices from age 16 to 18, most commonly in the form of 3 Advanced Level qualifications (A levels). Finally, again unlike the US, students wanting to

study for Bachelor's degrees then have to choose both an institution and subject (degree) at application stage. Such early subject specialisation, which begins at age 13/14, may be conducive to mismatch.

## 2.2 The datasets

We use individual-level administrative data on the population of state-school students in England for a single cohort. Our focus is on the cohort of young people who took their compulsory age 16 exams in 2006 and their non-compulsory exams two years later in 2008, at the end of secondary school. The grades from these exams are used to determine which university a student will be admitted to. The students enter university in the autumn of that year at age 18 (the traditional age for university entry in England) or 19 if they took a gap-year (around 25 % of our sample - see Table 1). Our data cover students in all publicly funded English schools,<sup>6</sup> and we combine this with information on the university degree attended by these students anywhere within the UK (England, Scotland, Wales, and Northern Ireland). Finally, we also incorporate aggregated data on the earnings outcomes of an earlier university cohort, which are based on tax records. These datasets are described in more detail below.

Our schools data come from the National Pupil Database (NPD), and include basic demographic information (gender, ethnicity, English as an additional language, special educational needs) alongside exam results at ages 11, 16, and 18. There is substantial attrition over this period of education in the English system, since many pupils leave at the end of compulsory education, after exams at age 16 (around 60% of our cohort), and a smaller group leave at age 18 without going on to university (around 15% of our cohort). Our main interest is in the subgroup who go on to university, but we use information on the complete population of age 16 students to construct key variables, as we describe below. Starting with a population of around 590,000 pupils in the 2006 cohort, we initially restrict the sample to all university students who went to a state-school, and on whom we have information on exam results at age 18, which results in a final sample of 138,969.

Our linked data on degree attended<sup>7</sup> come from the Higher Education Statistics Agency (HESA). We use university entry information from 2008 and 2009, since a quarter of students in England delay university entry for one year after age 18 examinations. These data contain

---

<sup>6</sup> 93 percent of students attend publicly funded secondary schools in England (Table 2A, DfE, 2010)

<sup>7</sup> Note that as in Dillon and Smith (2017) we observe a collapsed version of the student-degree match process, in that we only observe the degree that they attend, rather than where they apply.

information on every student's degree in every higher education establishment in the UK. Our main estimates use a 23 subject classification to distinguish degrees within universities for both achievement- and earnings-based match. This classification distinguishes “Medicine & Dentistry” from “Nursing”, and “Economics” is separately classified from other Social Science disciplines.<sup>8</sup> We also have access to a more detailed, 631 subject classification for academic-based match, which we use in robustness checks below.

Our aggregated earnings data come from the new Longitudinal Education Outcomes (LEO) dataset, which are compiled from tax records by Her Majesties Revenue and Customs (HMRC) in the UK. We use the median earnings outcomes five years after graduation for the earliest available cohort, which is those who completed undergraduate degrees in 2009.<sup>9</sup> These data are available for all 23 subject categories at each university where a subject is offered.

## 2.3 Measuring socio-economic status

To construct a measure of students' socio-economic status we follow Chowdry et al (2013). We use information on whether a student was eligible for free school meals at age 16, alongside a set of variables which describe the neighbourhood in which they live at that age. The free school meals indicator is essentially an indicator of whether a student is from a household in receipt of state benefits (around 15 percent of students). We additionally include a set of neighbourhood characteristics taken from the 2001 Census. These measures are available at the Lower Super Output Area level, which is a neighbourhood containing around 700 households or around 1,500 individuals. These measures includes the proportion of individuals in the neighbourhood that: 1) work in managerial or professional occupations; 2) hold an A-level equivalent qualification or above; and 3) own their home. In addition, we also use the derived ONS Area Classifications (2001) and the 2007 Index of Multiple Deprivation.<sup>10</sup>

---

<sup>8</sup> The 23 subjects are: “Agriculture & Related Subjects”, “Architecture, Building & Planning”, “Biological Sciences (excluding Psychology)”, “Business & Administrative Studies”, “Combined”, “Computer Science”, “Creative Arts & Design”, “Economics”, “Education”, “Engineering & Technology”, “English Studies”, “Historical & Philosophical Studies”, “Languages (excluding English Studies)”, “Law”, “Mass Communications & Documentation”, “Mathematical Sciences”, “Medicine & Dentistry”, “Nursing”, “Physical Sciences”, “Psychology”, “Social Studies (excluding Economics)”, “Subjects Allied to Medicine (excluding Nursing)”, and “Veterinary Science”.

<sup>9</sup> A comparison of the ranking of earnings by institution and subject when individuals are age 29 illustrates that the ranking of courses are broadly stable at later ages (Belfield et al., 2018).

<sup>10</sup> The ONS Area Classification aggregates local demographic and socio-economic statistics from the 2001 Census to classify areas into 53 different “types”. The Index of Multiple Deprivation ranks all lower-layer super output areas in England from least to most deprived, based on income, employment, education, health, crime, barriers to housing and services, and living environment.

We combine these measures using principle components analysis to create a standardised index.<sup>11</sup> We use the whole population of state-school students at age 16 in the relevant cohort to construct the index, so throughout this paper “SES” refers to socio-economic position relative to the whole school-cohort population rather than relative to the university-attending sub-population. The final row of Table 1 illustrates that this results in 8 percent of our university-attending sample coming from the most disadvantaged families, and 34 percent from the least disadvantaged families.

Table 1 highlights the key characteristics of our sample by SES quintile. Women are overrepresented in higher education, making up 56 percent of the sample. A quarter of our sample took a gap year, with the least deprived families more likely to take a year out than the most deprived. There are only a small proportion of people with special educational needs in our sample as might be expected, with 5 percent of the most deprived families and 3 percent of the least deprived families being categorized in this way. Finally, there is a strong association between having English as an additional language, ethnic minority status, and low SES, with these groups accounting for a larger proportion of low SES families.

**Table 1: Summary Statistics**

	Quintile of SES					Gender		Total
	1st	2nd	3rd	4th	5th	Men	Women	
Personal characteristics								
Ethnic minority	0.38 (0.48)	0.31 (0.46)	0.21 (0.41)	0.11 (0.32)	0.09 (0.29)	0.17 (0.17)	0.18 (0.18)	0.17 (0.38)
English as an Additional Language	0.27 (0.45)	0.22 (0.41)	0.14 (0.35)	0.07 (0.26)	0.06 (0.23)	0.12 (0.12)	0.12 (0.12)	0.12 (0.32)
Special Educational Needs	0.05 (0.23)	0.05 (0.21)	0.04 (0.19)	0.03 (0.18)	0.03 (0.18)	0.05 (0.05)	0.03 (0.03)	0.04 (0.19)
Gap year	0.22 (0.41)	0.23 (0.42)	0.24 (0.42)	0.25 (0.43)	0.28 (0.45)	0.26 (0.26)	0.25 (0.25)	0.25 (0.43)
A*-C in EBACCs	0.23 (0.42)	0.32 (0.47)	0.39 (0.49)	0.46 (0.50)	0.55 (0.50)	0.42 (0.42)	0.45 (0.45)	0.44 (0.50)
Proportion of sample	0.08	0.14	0.19	0.24	0.34	0.44	0.56	1.00
n	11697	19846	26468	33413	47084	61348	77621	138969

Source: NPD-HESA. n=138,969. Notes: A\*-C in EBACCs measures the percentage of students who achieve five or more grades A\* to C in traditional academic GCSE subjects (English, Maths, Science, Geography or History, and a language). Quintile of SES is defined out of the entire age 16 student population.

## 2.4 Two measures of student-degree match

We are interested in the match between student quality and degree quality. We calculate student quality according to age 18 exam test scores. Note our intention is to measure student

<sup>11</sup> See Appendix Figure A1 for a comparison of this measure to an alternative measure of parental socio-economic status from a linked data source. Results are comparable when using this alternative measure to capture socio-economic status, or the free school meals indicator alone, or a measure of parental education.

qualifications, rather than measure student ability, as student qualifications are the principal metric of which course the student can access.<sup>12</sup> We have two measures of degree quality, one based on the achievement of students on each degree, and one based on graduate earnings of previous cohorts of students on the degree, giving rise to two measures of student-degree match.

Each measure is calculated in three steps:

- (1) Calculate student quality: we rank individuals in the distribution of age 18 exam test scores based on their performance in their best three exams.<sup>13</sup>
- (2) Calculate degree quality: we rank each university-degree combination in a distribution of degree quality, based on either
  - (i) The median of the best three age 18 exam results of students on the degree (academic-based), or
  - (ii) The median earnings outcomes of an earlier cohort of students on the subject 5 years after graduation (earnings-based).

As mentioned in section 2.1, a distinctive feature of the UK education system is the importance of subject choices made in secondary education and at university. Our measure of individual and degree quality are based on the best three exam results. A levels are graded on a scale of A/B/C/D/E which are worth 270/240/210/180/150 QCA (Qualifications and Curriculum Authority) points respectively. Students typically study three A levels in different subjects, and the majority of universities set their entry requirements according to this measure. However a further complication is that some subjects are considered by universities to be more rigorous than others. This can be explicit, for example by naming ‘facilitating’ or ‘preferred’ subjects, and other times implicit in the offers that universities make to potential students (Dilnot, 2018).

To account for these differences in universities’ subject preferences, we follow Kelly (1976) and Coe et al. (2008) in calculating a subject difficulty adjustment, using an iterative approach based on our samples’ performance in different combinations of age 18 exams. For example, if students who took the same set of subjects consistently scored higher in one of these subjects, that subject would be deemed easier and would be awarded less points. This is

---

<sup>12</sup> In our conditional estimates of the match gap we control for student ability using the complete set of test scores from compulsory national examinations at age 11 and 16.

<sup>13</sup> We consider only the students who go on to university, so the relevant exam results distribution is that of university attendees. Some students take degrees that are equivalents to A-Levels. In these cases we calculate their A-Level equivalence scores.

iterated over all students and subject groupings until the difficulty adjusted scores are equalised. This is explained in more detail in Appendix. Figure A2 illustrates the difficulty ratings calculated for each subject, with the most difficult subjects being mathematics and natural sciences. We use these difficulty adjusted points when ranking students and degrees<sup>14</sup>.

As a final step we:

- (3) Calculate match: We subtract the student's percentile in the exam results distribution from the percentile of their degree on the quality distribution.

We therefore have two continuous measures of match for each student, an academic-based measure and an earnings-based measure. The continuous nature of our outcomes allow us to analyse inequalities across the severity of mismatch, rather than relying on arbitrary thresholds to categorize students as matched or not. Both measures represent the distance of each student's chosen degree from their position in the achievement distribution. With both measures of match, a student at the 50<sup>th</sup> percentile of the A-level distribution would be considered matched if they are enrolled on a degree at the median of the quality distribution. If a student attends a degree at a lower percentile than their own percentile in the student quality distribution, we consider them undermatched. If they attend a degree which ranks above their position in the student quality distribution, we consider them overmatched.

The academic-based measure of match measures whether students are enrolling in the degrees of the level of academic prestige that one might expect, given their qualifications. The earnings-based measure of match measures whether students are enrolling in degrees with the level of earnings that we might expect, given their qualifications. The latter is in the spirit of a classical human capital assumption, namely that students should expect earnings outcomes which are broadly comparable with their place in the achievement distribution. But it also has implications for social mobility, if low SES students are found to choose degrees with lower potential returns.

Both measures reflect different aspects of degree quality, and the same degree can be at quite different relative positions. For example, a degree which is positioned near the bottom of the achievement distribution, Computer Science at Southampton Solent, is considered high quality in terms of earnings, ranked at the 70<sup>th</sup> percentile. In contrast, English at Edinburgh is

---

<sup>14</sup> A parallel set of results using the un-diffculty adjusted rankings are also available in Appendix Table A1. All results are qualitatively similar.

ranked at the 90<sup>th</sup> percentile on our academic-based quality measure, but is only ranked at the 35<sup>th</sup> percentile in terms of our earnings-based measure. The correlation between the two measures of 0.58. Appendix Figure 3 presents a scatterplot of each course by these degree quality measures.

## 2.5 Methods

To understand the nature of student matching we use three distinct methods to present the results. First, we show a simple plot of students' achievement decile against average degree quality for all students in that decile. If all students were perfectly matched to their degrees this line would be straight and at a 45-degree angle.<sup>15</sup> The extent to which a point is above a 45-degree line indicates how overmatched these students are on average, and similarly the distance below the 45-degree line reveals the extent of undermatch. It imposes minimal assumptions beyond those involved in the creation of the metrics. Plotting this match-line for different types of students allows us to study the match gap at any point in the achievement distribution.

Second, we estimate SES and gender gaps in match, conditional on individual characteristics and achievement prior to age 18. Specifically, we estimate the following regressions:

$$M_{ia} = \beta_0 + \sum_{j=2}^5 \beta^j I(SES_i = j) + \gamma female_{ia} + \delta X_{ia} + \pi P_{ia} + \varepsilon_{ia}, \Delta a \quad (1)$$

Where  $M_i$  is our measure of match,  $\widehat{\beta^j}$  represents our estimated SES gap in match, and  $\hat{\gamma}$  is our estimated gender gap in match, conditional on background characteristics ( $X_i$ ) and prior achievement at age 11 and 16 ( $P_i$ ). Given that achievement is used to define match, there will be ceiling and floor effects; it would be impossible for the lowest ranked students to undermatch or the highest rank student to overmatch. We therefore estimate the models separately, first across deciles of achievement ( $a$ ), before focusing on those in the top and bottom quintiles of achievement. Standard errors are clustered at the secondary school level.

While this approach estimates SES and gender gaps at the mean of our match outcomes, we are also interested in the size of these gaps across the distribution of match: in particular the gaps for those who severely under or overmatch. It could be the case that our OLS estimates mask much larger SES and gender inequalities in match in the tails of the distribution. A key

---

<sup>15</sup> This is feasible, even if at the extremes of the quality distribution it would require there to be no variation in student quality within a degree. For example the courses at the top percentile of the quality distribution would need all their students to be from the top percentile. There is evidence of near perfect matching in our data, for the top ranked degree of Maths at Cambridge more than half of the 92 students enrolled have 'best 3' A-grades in either Physics, Maths, Further Maths, or in Chemistry, Physics, Further Maths.

strength of our approach is that our continuous measure of match allows us to consider this for the first time. To this end, our third approach is to use unconditional quantile regression (UQR) to estimate SES and gender match across the distribution of match<sup>16</sup>. We use a Re-centred Influence Function (RIF) regression (Firpo et al., 2009), specifying our distributional statistic of interest as the quantiles of our match variable  $q_\tau$  where  $\tau$  is each decile from 1 to 9.

$$RIF(M_{ia}; q_\tau) = \beta_0^\tau + \sum_{j=2}^5 \beta^j I(SES_i = j) + \gamma^\tau female_{ia} + \delta^\tau X_{ia} + \pi^\tau P_{ia} + \varepsilon_{ia}, \Delta a \quad (2)$$

Here, our coefficients  $\widehat{\beta}^\tau$  and  $\widehat{\gamma}^\tau$  illustrate the estimated SES and gender inequalities in match. Given that we estimate our models by achievement quintiles ( $a$ ), for high-attainers this will estimate the SES and gender gaps from the most severely undermatched (10<sup>th</sup> percentile) to those who are matched (90<sup>th</sup> percentile). For low-attainers, this will estimate the SES and gender gaps for those who are matched (10<sup>th</sup> percentile) up to the most severely overmatched (90<sup>th</sup> percentile).

Finally, to explore the potential drivers of SES and gender gaps in match, we consider a range of different mechanisms by including a series of separate (bad) controls  $Y_{ia}$  in model (1). Here we are interested in how much these reduce our estimated SES and gender gaps.

$$M_{ia} = \beta_0 + \sum_{j=2}^5 \beta^j I(SES_i = j) + \gamma female_{ia} + \vartheta Y_{ia} + \delta X_{ia} + \pi P_{ia} + \varepsilon_{ia}, \Delta a \quad (3)$$

To explore the role of geography ( $Y_{ia} = \varphi dist_{ia}$ ), we show three alternative specifications: one controlling for distance to university attended in kilometres, one controlling for distance to each of the nearest three universities to the student's home neighbourhood, along with the distance to all remaining universities (similar to Gibbons and Vignoles, 2012), and one controlling for distance to the nearest matched university. To disentangle subject choice at university from the choice of institution, we include 23 subject categories in model (3) (where  $Y_{ia} = \sum_{k=2}^{23} \sigma^k I(Subject_i = k)$ ). This will account for the average mismatch of students studying a certain subject area e.g. students studying history being earnings undermatched, so that any remaining SES and gender gaps in this specification can be interpreted as likely institutional-driven inequalities, within subject of study. Finally, again in separate regressions, to explore the role of school-level factors in driving SES and gender inequalities in match, we control for school fixed effects ( $Y_{ia} = \omega school_{ias}$ ). To investigate the role of possible school

---

<sup>16</sup> We use UQR rather than the standard conditional quantile regression as we want to estimate these inequalities at given points in the unconditional distribution of match, rather than the residual match distribution, after the confounders in model (1) are accounted for.



mechanisms we include a range of observable school factors, including the achievement of the school, the provision and uptake of ‘facilitating’ or ‘preferred’ subjects in the school, the proportion of students from the school that attend university and high-status universities, and the proportion of students who are high SES at the secondary school attended.

### 3. Results

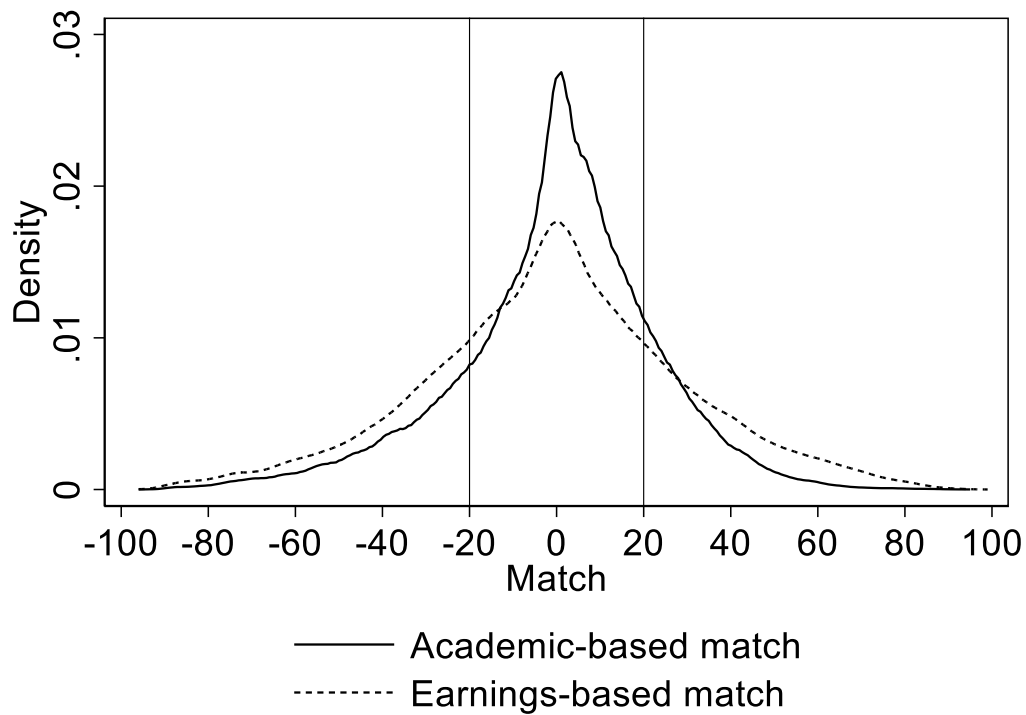
Figure 1 shows the distribution of the two measures of student-degree match which result from this process. Both measures have peaks with students being well matched and are approximately symmetrical. The earnings-based measure is more dispersed than the academic-based measure. This reflects that there are observed academic-based entry requirements for enrolling on a degree. There are no such restrictions in terms of later earnings, and students are likely to be less well informed of the potential earnings of each degree.

Using the binary definition of mismatch from Dillon and Smith, 2017 where mismatch of is +/- 20 percentiles from the matched degree, 16 percent of our sample are overmatched and 16 percent of our sample are undermatched using our academic-based measure. For our earnings-based measure 22 percent overmatch and 23 percent undermatch. Dillon and Smith (2017) find around 25% of students in the US are overmatched and 25% undermatched according to their composite college-input-quality measure<sup>17</sup>. This is most comparable to our academic-based measure of match, and while it would be problematic to draw strong conclusions from this, the comparison is suggestive that there is more mismatch in the US than in the UK. Figure 2 further highlights the strength of our approach in being able to analyse the extent of inequalities in mismatch in the tails of the distribution. 3% of our sample are undermatched by over 50 percentiles using our academic-based measure, and 5% are undermatched by over 50 percentiles using our earnings-based measure. 1% overmatch by more than 50 percentiles using our academic-based measure, and 5% overmatch by more than 50 percentiles using our earnings-based measure. In section 3.3 we will explore how SES and gender gaps vary for the most severely under- and over-matched.

---

<sup>17</sup> Dillon and Smith’s college quality measure comprises 4 measures of quality – the mean SAT score (or ACT score converted to the SAT scale) of entering students, the percent of applicants rejected, the average salary of all faculty engaged in instruction, and the undergraduate faculty-student ratio.

**Figure 1: Academic-based and earnings-based measures of student-degree match**



Source: NPD-HESA.  $n=138,969$ . Notes: Academic-based match defined by degrees' median student age 18 achievement percentile minus student's age 18 achievement percentile. Earnings-based match defined by degrees' median graduate earnings percentile minus student's age 18 achievement percentile.

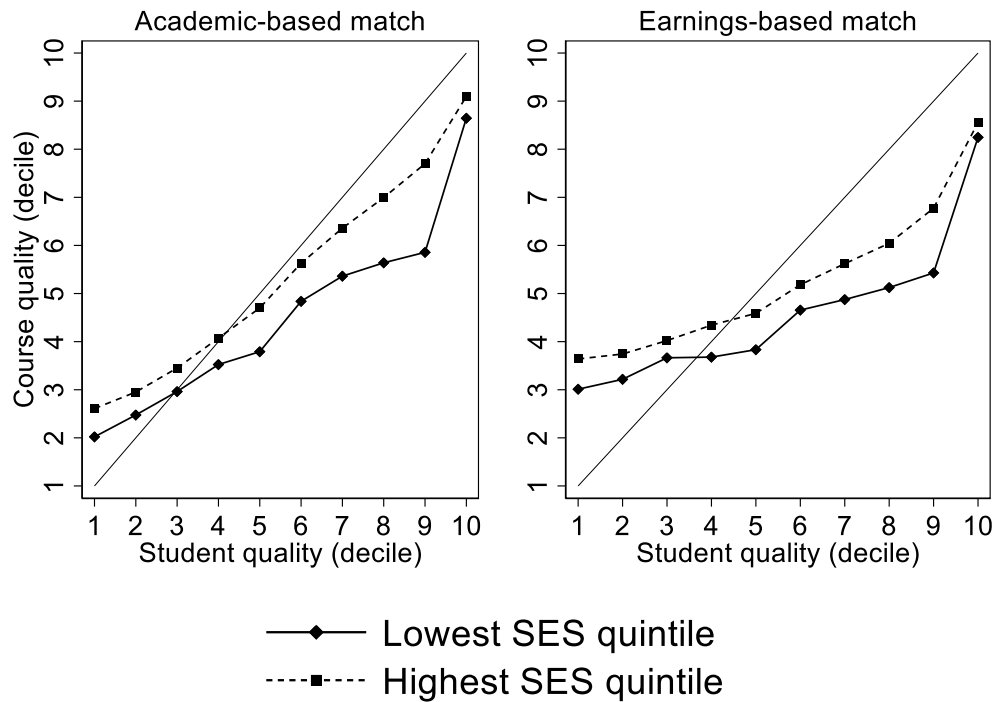
### 3.1 Inequalities in Match Gaps by Achievement

Figure 2 plots our two degree quality measures against student achievement for high and low SES pupils, illustrating raw gaps in points- and earnings-based match by SES across the achievement distribution (left and right panel respectively). For both match measures we see that the relationship is approximately linear and is flatter than 45 degrees, meaning that low-attainers are more likely to overmatch and high-attainers are more likely to undermatch (as previously described, reflecting floor and ceiling effects). As would be expected given the distributions in Figure 2, we see that the earnings match curve is flatter than the points match, meaning that there is more mismatch in terms of earnings than points.

For both measures we see stark SES gaps in match. For every given percentile of individual achievement, high SES pupils attend higher ranked degrees than low SES pupils. The SES match gap increases in the top half of the distribution of student achievement, with the exception of the top decile of students where the gap is the smallest. As much of the previous literature on mismatch has focused on high-attaining low SES students (Hoxby and Avery, 2012; Black, Cortes and Lincove, 2015), this implies they may be underestimating the

extent of mismatch, by failing to study those students for whom mismatch is largest, between the 70<sup>th</sup> and 90<sup>th</sup> percentiles of achievement. These patterns hold for both the points- and earnings-based match measures.

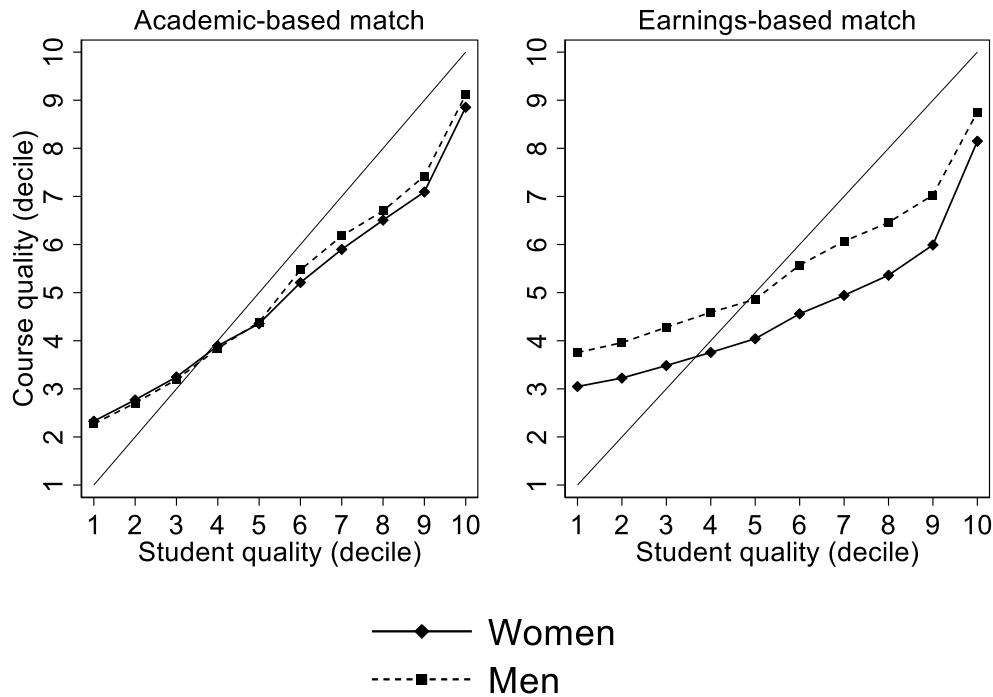
**Figure 2: SES match by student achievement**



Source: NPD-HESA. n=138,969. Notes: The 45-degree line represents perfect matching throughout the achievement distribution. Academic-based match defined by degrees' median student age 18 achievement percentile minus student's age 18 achievement percentile. Earnings-based match defined by degrees' median graduate earnings percentile minus student's achievement percentile. Student quality defined as their age 18 achievement decile.

Figure 3 next plots gender gaps in match for our two degree quality measures. Unlike our findings for SES, the findings differ across measures of match. For our academic-based match we observe almost no gender gap in match. In the bottom half of the achievement distribution men and women attend degrees that are equally academically selective, and in the top half men are enrolling in degrees with slightly higher peer achievement. By contrast, the earnings match measure highlights striking gender gaps. Men consistently attend degrees with graduate earnings around one decile higher than women across the distribution of achievement. This gender gap narrows in the top achievement decile, but even then males with the same subject difficulty-adjusted achievement are still enrolling in degrees with higher median earnings. In the next section, we test the robustness of these gaps at the top and the bottom of the achievement distribution, by conditioning on characteristics and prior academic achievement.

**Figure 3: Gender match by student achievement**

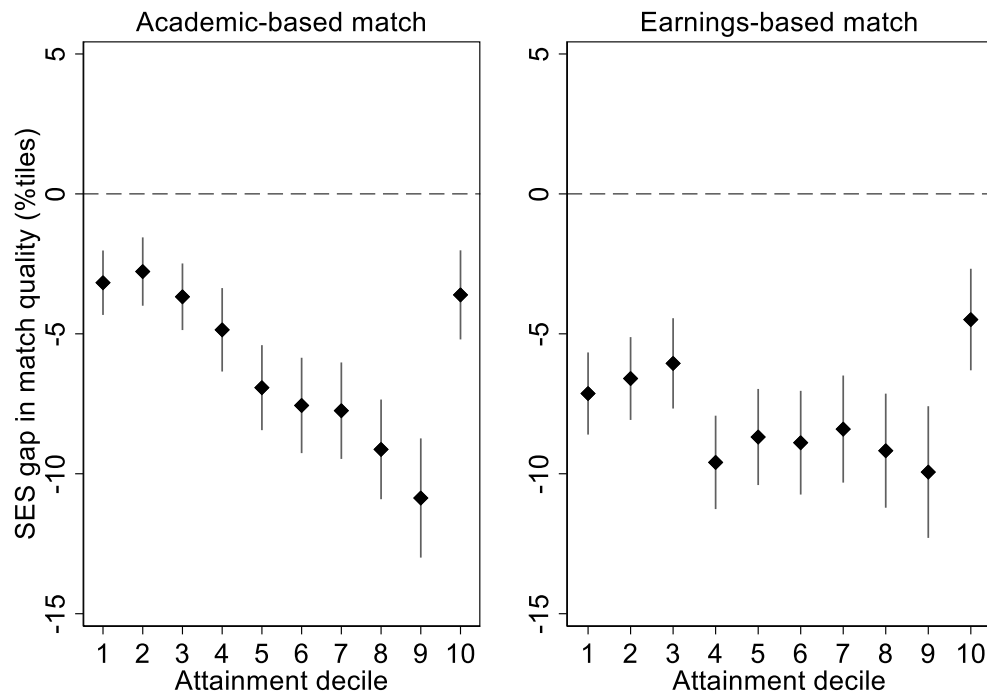


Source: NPD-HESA. n=138,969. Notes: The 45-degree line represents perfect matching throughout the achievement distribution. Academic-based match defined by degrees' median student age 18 achievement percentile minus student's age 18 achievement percentile. Earnings-based match defined by degrees' median graduate earnings percentile minus student's achievement percentile. Student quality defined as their age 18 achievement decile.

### 3.2 Conditional Match Gaps

Figures 4 and 5 present estimates of the match gaps conditional on student characteristics and prior achievement up until age 16 (equation 1) across the distribution of achievement. Each point represents a separate regression for each achievement decile. Figure 4 plots the achievement gap between the lowest and highest SES quintile, and shows that conditional on demographics and prior achievement, the SES gap is increasing across the achievement distribution up to the ninth decile of achievement, where low SES students undermatch by 11 (10) percentiles more than high SES students for our academic- (earnings-) based measure of match. For top performing students the SES match gap reduces significantly to 4 percentiles. This implies that there are factors in play that ensure that the very best students are well matched to degrees regardless of their level of disadvantage. The largest SES match gaps are found for the above average students, a group of students that have largely been passed over by the literature.

**Figure 4:SES conditional match inequalities**

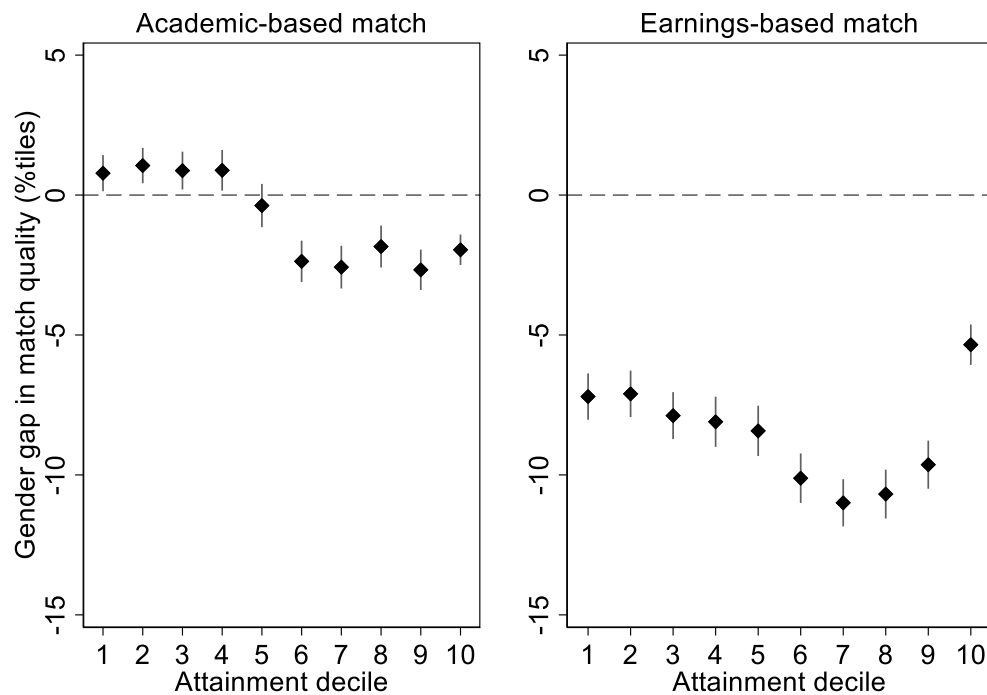


Source: NPD-HESA. n=138,969. Notes: Each point represents the SES match gap between groups 1 and 5 from specification 1, estimated for each decile of the student achievement distribution. Controls include dummies for ethnicity, English as an Additional Language, Special Educational Needs, and gap year before college, and cubics in age 11 and age 16 exam results. We present the 95 percent confidence intervals, with standard errors clustered at the secondary school level.

Figure 5 shows a small conditional gradient in undermatch for high-attaining women, relative to men for academic-based match, with a more pronounced conditional gender gap across the entire distribution of achievement for our earnings-based match measure. Again, the top decile of attainers shows a smaller gender gap for this match measure, indicating that the highest attaining women are more similarly matched, relative to the highest attaining men.

Having explored the inequalities in match across the entire distribution, for the remainder of the paper we focus on those students in the top and bottom quintiles of achievement for brevity. Table 2 presents conditional estimates replicating Figures 4 and 5 for these quintiles. Each of the four columns represents a separate regression. The SES parameters represent the match gaps for each SES quintile relative to the highest.

**Figure 5: Gender conditional match inequalities**



Source: NPD-HESA. n=138,969. Notes: Each point represents the gender match gap from specification 1, estimated for each decile of the student achievement distribution. Controls are dummies for ethnicity, English as an Additional Language, Special Educational Needs, and gap year before college, and cubics in age 11 and age 16 exam results. We present the 95 percent confidence intervals, with standard errors clustered at the secondary school level.

**Table 2: SES and gender conditional match gaps**

Achievement Quintile SES Quintile	Academic-based match		Earnings-based match	
	(1) Lowest (1 <sup>st</sup> )	(2) Highest (5 <sup>th</sup> )	(3) Lowest (1 <sup>st</sup> )	(4) Highest (5 <sup>th</sup> )
1st	-2.53 (0.46)***	-8.33 (0.81)***	-6.29 (0.57)***	-8.25 (0.88)***
2nd	-2.42 (0.39)***	-4.47 (0.47)***	-3.07 (0.48)***	-4.45 (0.53)***
3rd	-1.69 (0.34)***	-3.29 (0.34)***	-1.72 (0.44)***	-3.89 (0.44)***
4th	-0.62 (0.33)	-1.83 (0.27)***	-1.28 (0.44)**	-2.27 (0.36)***
Women	0.69 (0.24)**	-2.44 (0.25)***	-7.48 (0.32)***	-8.07 (0.32)***
Constant	14.35 (0.33)***	-17.88 (0.57)***	28.15 (0.47)***	-21.26 (0.56)***
Clusters	2135	2005	2135	2005
n	27794	27786	27794	27786
Controls	Yes	Yes	Yes	Yes

Source: NPD-HESA. n=138,969. Controls are dummies for ethnicity, English as an Additional Language, Special Educational Needs, and gap year before college, and cubics in age 11 and age 16 exam results. \*\*\* Significant at the 1% level, \*\* Significant at the 5% level, \* Significant at the 10% level.

The positive constant for low-attainers indicates that they overmatch on average while the negative constant for high-attainers indicates that they undermatch on average as shown in Figures 2 and 3. Therefore, the coefficients in columns 1 and 3 represent the SES or gender gap in overmatch, and the coefficients in columns 2 and 4 represent the SES or gender gap in undermatch.

The results in column 2 are similar to previous findings on mismatch that consider the extent of academic undermatch among high-attaining students. We find that there is an 8.3 percentile gap in match for those from the lowest SES quintile relative to those from the highest SES quintile using our academic-based measure of match. This is consistent with the findings in the literature (e.g Hoxby and Avery, 2012; Smith, Pender and Howell, 2013) that high-attaining disadvantaged pupils are more likely to undermatch than their more advantaged counterparts. This 8.3 percentile gap corresponds to the difference between studying economics at the London School of Economics (ranked 5th in the Times Higher UK university rankings) versus Exeter (ranked 18th). This could have real labour market consequences for the student; the median earnings difference five years after graduation between these two degrees is £13,200 per year. The extent of the match gap closes as the difference in the SES quintile narrows: the points gap between the highest SES quintile and the second, third and fourth quintiles are 4.5, 3.3 and 1.8 percentiles respectively.

Column 1 re-estimates these gaps for students from the lowest achievement quintile. Despite all students being in the lowest 20% in terms of achievement, the high SES students attend degrees with higher attaining peers. The mismatch gap is 2.5 percentiles, implying that low SES students overmatch by 2.5 percentiles less than high SES students. This SES match gap is about a quarter of the size of that for high attaining students.

Columns 3 and 4 estimate the earning-based match gaps. For high attaining students (column 4) the SES earnings-based match gaps are of the same magnitude as the academic-based gaps. However, for low attaining students the SES earnings-based gap is three times larger than the academic-based gap. Low-attaining low SES pupils undermatch in earnings by 6 percentiles more than their low-attaining high SES counterparts. This suggests that even among low attainers, more advantaged pupils are more likely to attend degrees with higher labour market rewards. These results represent the first key finding of the paper; that low SES students undermatch more, and overmatch less than high SES students, and this is true for both academic-based match, and earnings match.

The remaining parameter of interest is the coefficient for the female indicator variable, which shows the gender match gap. Conditional on student characteristics and prior achievement we find no significant difference among low-attaining students in terms of academic-based match. For high-attaining students, women undermatch by 2.4 percentiles more than men (columns 1 & 2). In contrast, the gender gap is large when considering the earnings-based match in columns 3 & 4. These gaps are of a similar magnitude to the SES gaps, with both low- and high- attaining women undermatching by 7-8 percentiles more than men. This is in line with the raw plots of enrolment by student achievement seen in Figure 4. This is the second of our key findings; that while women attend degrees that are almost as academically selective as men, at every point on the achievement distribution they attend degrees with substantially lower rewards on the labour market. We will return to potential drivers of these gaps, including preferences, in section 4.

### 3.3 Severity of match

While Table 2 illustrates SES and gender gaps in match for the mean level of match, it may be the case that these inequalities vary across the distribution of match. In particular, we are interested in the extent of these inequalities among cases where students are severely under or overmatched. Figures 6 and 7 explore this using unconditional quantile regression (equation 2) for our earnings-based match measure (see Appendix Figures A4 and A5 for academic-based match). We plot the SES and gender gaps from the 10<sup>th</sup> to the 90<sup>th</sup> percentile of the distribution of match for low- and high-attainers. Recall that low-attainers typically overmatch, whereas high-attainers typically undermatch. Therefore, for low attainers (left hand panel of figures 6 and 7), the x-axis runs from those who are matched (at the 10<sup>th</sup> percentile) to those who are severely overmatched (90<sup>th</sup> percentile). For high-attainers (right hand panels of figures 6 and 7) the x-axis runs from those who are severely undermatched (10<sup>th</sup> percentile) to matched (90<sup>th</sup> percentile). In each case, the estimates represent the earnings-based match gap between the top and bottom SES quintile or between genders. A negative value in Figure 6 represents the degree to which those from the lowest SES quintile are less overmatched (low-attainers), or more undermatched (high-attainers), compared with those from the top quintile. Similarly for Figure 7, a negative value represents the degree to which females are less overmatched (low-attainers) or more undermatched (high-attainers) compared with males.

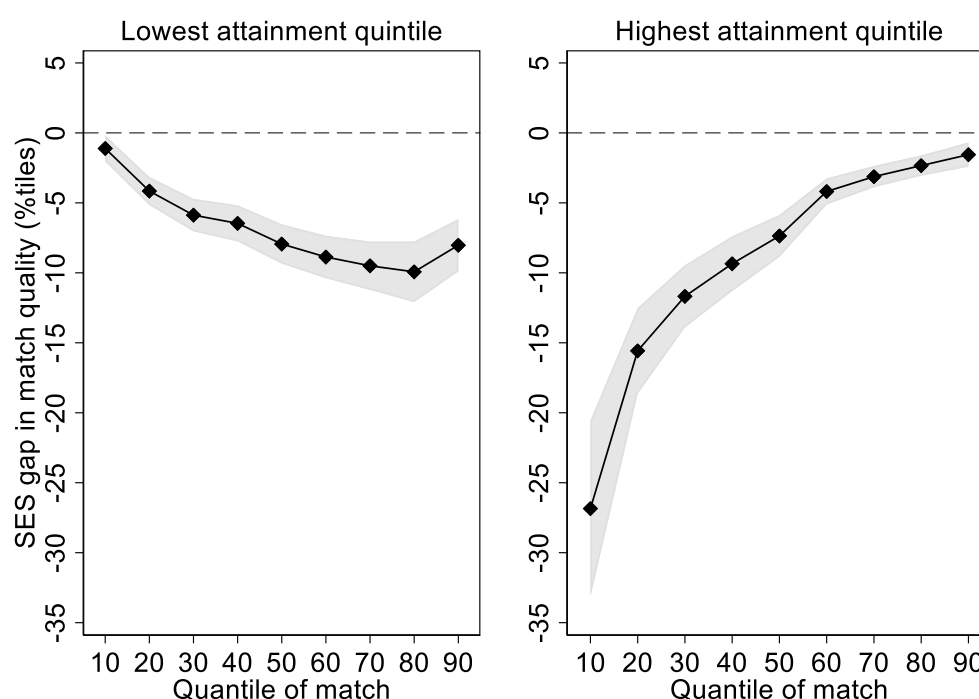
Looking first at low-attainers (left hand panel of Figure 6) we see that the SES gap is very small for students who are well matched (10<sup>th</sup> percentile). However as we move along the distribution from matched to severely overmatched, the gap becomes more pronounced. The



implication of this negative gradient is that even within the group of low-attaining students who manage to significantly overreach themselves in terms of the degree they eventually access, students from richer backgrounds still manage to reach further – attending higher earning degrees - than poorer students.<sup>18</sup>

A similar pattern is observed with the gender gap for low-attaining students (Figure 7, left hand panel), albeit that the gap is larger throughout the mismatch distribution. For well-matched students, females still attend degrees that are 3.3 percentiles less overmatched than males, and this gap increases to 11.6 percentiles for severely overmatched students.

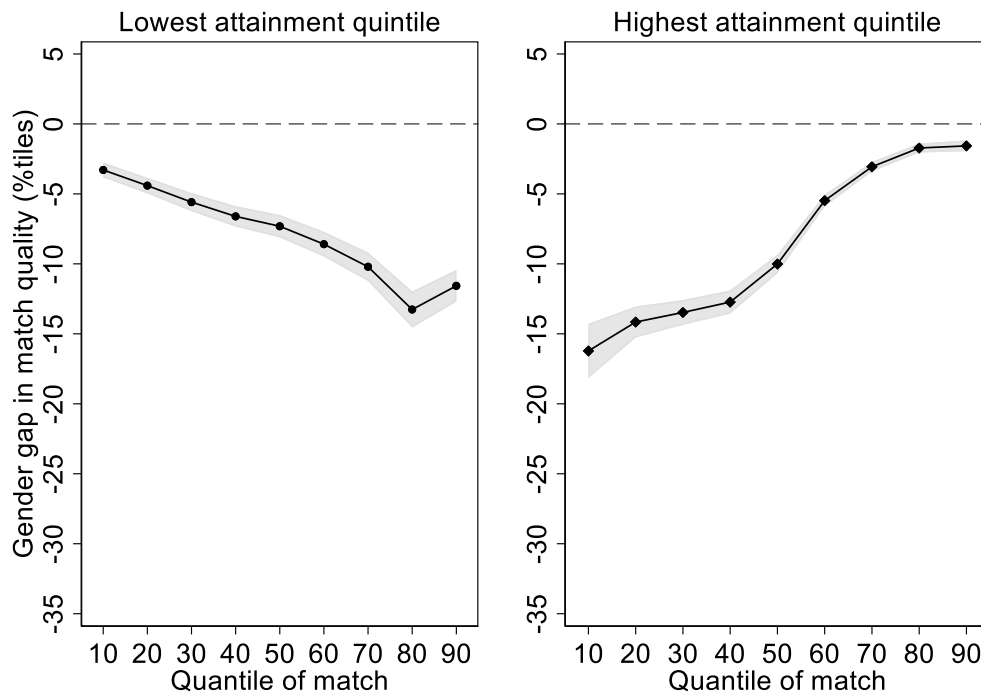
**Figure 6: SES gaps in severity of earnings-based match**



Source: NPD-HESA. n=138,969. Notes: Each point represents the SES match gap between groups 1 and 5 from specification 2, estimated for each decile of the match distribution. Controls are dummies for ethnicity, English as an Additional Language, Special Educational Needs, and gap year before college, and cubics in age 11 and age 16 exam results. We present the 95 percent confidence intervals, with standard errors clustered at the secondary school level.

<sup>18</sup> This only holds for our earnings measure of match. When we use our academic-based measure of match (see Appendix Figures A3 and A4), low-attaining low SES pupils overmatch to a similar extent to high SES pupils and women overmatch to a similar extent to men, when considering those who severely overmatch.

**Figure 7: Gender gaps in severity of earnings-based match**



Source: NPD-HESA. n=138,969. Notes: Each point represents the gender match gap from specification 2, estimated for each decile of the match distribution. Controls are dummies for ethnicity, English as an Additional Language, Special Educational Needs, and gap year, and cubics in age 11 and age 16 exam results. We present the 95 percent confidence intervals, with standard errors clustered at the secondary school level.

For high-attainers, (right hand panels of Figures 6 and 7), the situation is reversed. The SES and gender gaps are largest for the most severely undermatched students (at the 10<sup>th</sup> percentile of match). In addition to the positive gradient, the size of the gaps are larger for the high attaining students. For the most severely undermatched students, low SES students undermatch 27 percentiles more than high SES students, and women undermatch 16 percentiles more than men. This holds also for the academic-based match measure (see Appendix Figures 4 and 5), with the SES gap among high-attaining students being 32 percentiles and women undermatching more than men by 9 percentiles. This suggests that even among high-attainers who are severely undermatched, low SES students and women are attending degrees that attract far lower financial rewards<sup>19</sup> than they could, compared to those from richer backgrounds and men.

<sup>19</sup> And are attending degrees that are substantially less academically selective

#### 4. Robustness

The construction of our match measure requires us to make a number of decisions. In this section, we use the detailed and extensive nature of our dataset to test these robustness of our findings to these decisions by constructing alternative match measures. Table 3 presents estimates for our academic- and earnings-based measures of match for high- and low-attaining students, with the first two columns showing the baseline estimates from Table 2, and a number of alternative specifications in the remaining columns for comparison (Appendix Table A1 presents further robustness checks across additional specifications). Our aim is to demonstrate that our results are robust to a number of alternative model choices.

For our main measure of match, as described in Section 2, we adjust the points associated with each A-level grade to account for the difficulty rating of each subject. However we might be concerned if certain groups of students choose different types of subjects at A-Level, in which case the difficulty adjusted measure may be endogenous to student SES. An alternative method of dealing with the potential endogeneity of A level subject choice is to use earlier, broader measure of student achievement. In Columns 3-4 of Table 3 we therefore rank students based on their qualifications from compulsory education at age 16 (GCSE level). Typically students study 10 subjects at this level, and these qualifications are not the main feature of the university application process. We sum the scores across subjects for each student and then calculate their national percentile rank. As with our standard academic-based match measure, the degree ranks are calculated on the basis of the median student on each degree, replacing our standard measure of achievement with the scores from compulsory subjects at age 16.

In columns 5 and 6 we re-calculate the student and degree quality measures within university subject choice. For example, to calculate our academic-based match measure, for a student who is observed as studying nursing, we calculate their national academic-based percentile rank among all individuals are studying nursing (as opposed to all students). We then calculate the national percentile rank (in terms of academic achievement or earnings) of each nursing degree among all nursing degrees (as opposed to all degrees). This measure has the advantage that students enrolling in the same subject area are likely to have a more similar set of qualifications than students studying different subject areas, making the academic ranking more comparable. The disadvantage is that it implicitly assumes that students choose to study one subject and then choose across universities, rather than applying to study different subjects within the same university (or across universities). Note that the results here will express the

mismatch gap within subject, i.e. the gap that remains once subject choice has been taken into account.

Finally, our approach contrasts with much of the existing US literature in that we can observe match at the degree (subject\*institution) rather than institution level. In Columns 7-8, for comparability, we condense our data to create a more comparable measure of match, by measuring university quality according to the median student at each university.

In summary, adopting almost all of these alternative measures of match does not result in any substantial changes to our main findings – low SES students are more undermatched and less overmatched than high SES students in terms of both academic-based and earnings-based match, and high-attaining women are more undermatched than men in terms of earnings-based match. There is one exception: the gender gap in earnings match is substantially reduced (though not entirely eradicated for high attainers, falling from -8 to -1.9 percentiles) when we re-calculate match within subject. This implies that, conditional on subject chosen, women are not attending institutions with lower returns, but are consistently choosing subjects with lower earnings throughout the achievement distribution. We will return to this issue in detail in the following section.

**Table 3: SES and gender conditional match gaps across alternative specifications**

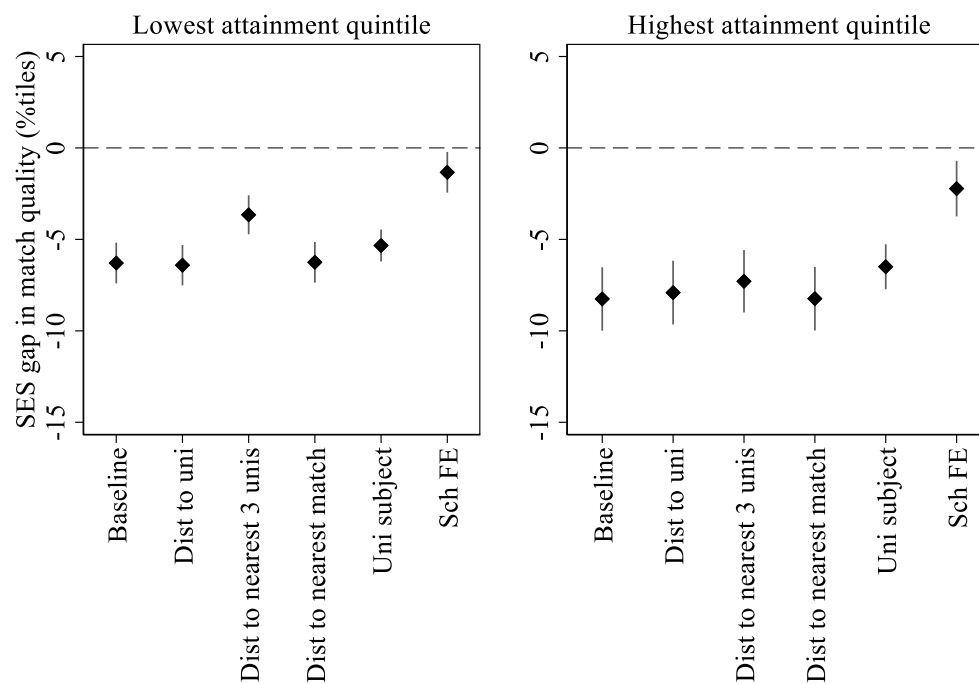
Achievement quintile	Academic-based match							
	Baseline		GCSE-based		Within subject		University level	
	Lowest	Highest	Lowest	Highest	Lowest	Highest	Lowest	Highest
SES quintile								
1 <sup>st</sup> (Lowest)	-2.53 (0.46)***	-8.33 (0.81)***	-3.43 (0.80)***	-7.70 (1.07)***	-3.76 (0.53)***	-7.85 (0.92)***	-2.57 (0.49)***	-9.08 (0.91)***
Women	0.69 (0.24)**	-2.44 (0.25)***	-3.25 (0.40)***	-5.57 (0.54)***	-0.74 (0.28)**	-2.54 (0.36)***	-2.40 (0.25)***	-3.79 (0.29)***
Clusters	2135	2005	2135	2005	2135	2005	2135	2005
N	27794	27786	27794	27786	27794	27786	27794	27786
Achievement quintile	Earnings-based match							
	Lowest	Highest	Lowest	Highest	Lowest	Highest	Lowest	Highest
	Lowest	Highest	Lowest	Highest	Lowest	Highest	Lowest	Highest
SES quintile								
1 <sup>st</sup> (Lowest)	-6.29 (0.57)***	-8.25 (0.88)***	-7.36 (0.94)***	-5.92 (1.11)***	-8.57 (0.69)***	-9.65 (0.98)***	-7.31 (0.60)***	-10.53 (0.96)***
Women	-7.48 (0.32)***	-8.07 (0.32)***	-10.00 (0.47)***	-10.98 (0.59)***	0.29 (0.31)	-1.86 (0.38)***	-1.99 (0.28)***	-3.93 (0.32)***
Clusters	2135	2005	2135	2005	2135	2005	2135	2005
N	27794	27786	27794	27786	27794	27786	27794	27786

Source: NPD-HESA. n=138,969. All specifications control for ethnicity, English as an Additional Language, Special Educational Needs, and gap year before college, and cubics in age 11 and age 16 exam results. The cubic in age 16 exam results is omitted from the GCSE-based regressions. SES baseline is category 5 (highest quintile) while the other quintiles are included in the model but not reported here.

## 5. Potential drivers

We now turn our attention to possible explanations for these SES and gender inequalities (again, for simplicity concentrating on earnings-based match)<sup>20</sup>, exploring three possible factors - subject choice, geography, and school attended. For each of these three potential sets of drivers, we condition on additional measures to investigate whether our SES and gender gradients in earnings-based match are reduced by the inclusion of these variables. Figures 8 and 9 present the SES and gender gap coefficients after each characteristic is separately added relative to the baseline conditional SES and gender gap reported from Table 2.

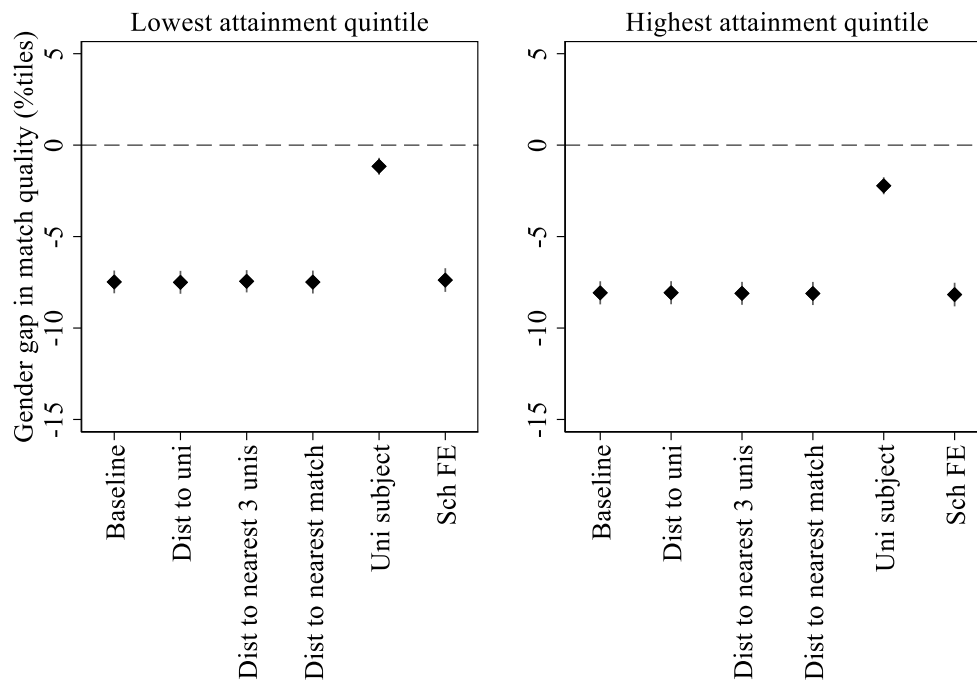
**Figure 8: SES gaps in earnings-based match, conditional on geography, subject, and schools**



Source: NPD-HESA. n=138,969. Notes: Each point represents the SES match gap between groups 1 and 5 from specification 3, estimated for the top and bottom quintiles of the achievement distribution. The baseline controls are dummies for ethnicity, English as an Additional Language, Special Educational Needs, and gap year before college, and cubics in age 11 and age 16 exam results. We present the 95 percent confidence intervals, with standard errors clustered at the secondary school level.

<sup>20</sup> See Appendix Figures A5 and A6 for academic-based match

**Figure 9: Gender gaps in earnings-based match, conditional on geography, subject, and schools**



Source: NPD-HESA. n=138,969. Notes: Each point represents the gender match gap from specification 3, estimated for the top and bottom quintiles of the achievement distribution. The baseline controls are dummies for ethnicity, English as an Additional Language, Special Educational Needs, and gap year before college, and cubics in age 11 and age 16 exam results. We present the 95 percent confidence intervals, with standard errors clustered at the secondary school level.

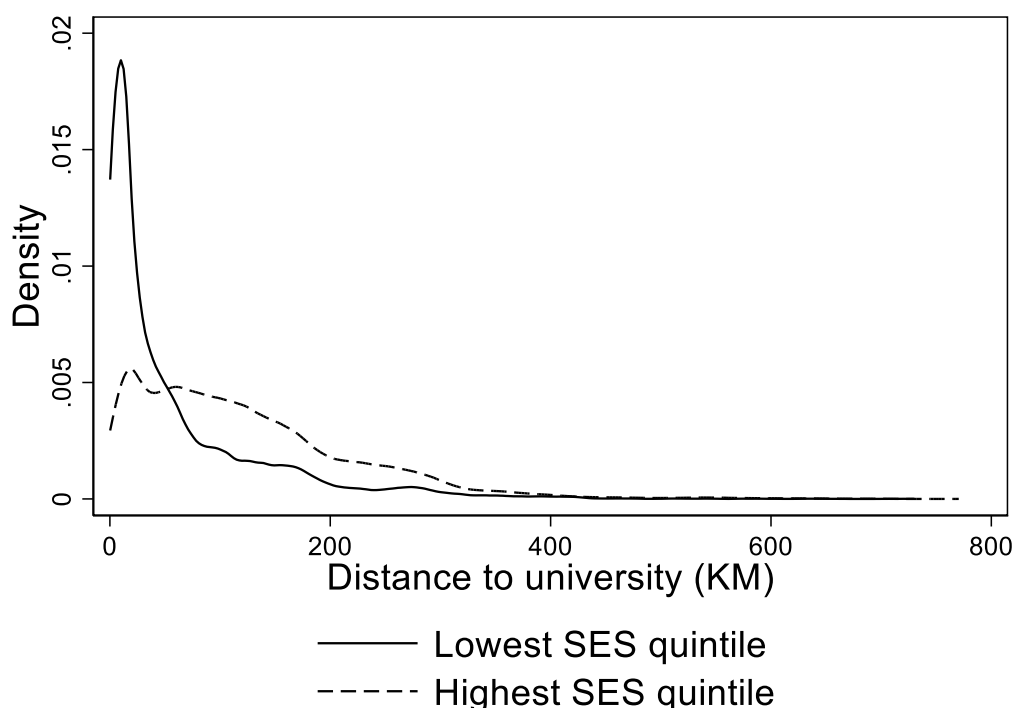
## 5.1 Geography

Geography is often highlighted as a key driver of match (Hoxby and Avery, 2012). A simple plot of distance to university attended by SES, as presented in Figure 10, shows that there is a substantial SES gap in distance travelled to university. In particular, low SES students are far more likely to be found at universities close to their home location.<sup>21</sup> If the SES gap in match is driven by geography, with high SES students travelling further in order to achieve a better match then conditioning on distance to university should reduce the gap. However, we find that the inclusion of distance to university attended has no impact on the SES gap for high- or low-

<sup>21</sup> Here, we define their home location using the centre of the student's neighbourhood, defined at the Lower Super Output Area level. This chart includes all students, but the results are similar if we restrict to high-attaining students only.

attaining students (third column in Figure 8).<sup>22</sup> Implying that low SES students undermatch to degrees regardless of distance. However, the distance to university attended is endogenous to the students' choice, therefore we test for the impact of distance using two other pre-determined geographic characteristics.

**Figure 10: Distribution of distance travelled to university for low and high SES students**



Source: NPD-HESA. n=138,969. Distance to university is calculated using the centre of the student's neighbourhood, defined at the Lower Super Output Area level. This chart includes all students.

The first pre-determined measure of proximity relates to the size of the choice set of universities close to the student's home. For this we calculate the total distance to the nearest three universities. If a student lives in an area with several institutions this should improve the probability of a good match, and if high SES students are more likely to be located in such areas this could contribute to the gap. We find this to be the case for low-attaining students, the inclusion of this term reduces the SES gap by almost half from 6 to 3.5 percentiles (see Figure

<sup>22</sup> Other measures of distance to university attended, such as log or quadratics specifications produce similar results.

8). In contrast the inclusion of this parameter has little impact on high-attaining students, suggesting these students are less reliant on universities in their local area.

Our second pre-determined measure of higher education proximity is the distance to a matched degree, where we define match as a student attending a degree whose quality percentile is within 20 percentage points of their achievement percentile. As with the distance to university attended with find that conditioning on geography in this dimension has little impact on the SES gap for either low or high attaining students (Figure 8). Therefore, we conclude that for students in England, distance to university does not differentially impact the match to degrees for high or low SES students. The same is true for the gender gap: geographical factors do not reduce the gender gap in match (Figure 9), suggesting that males are not attending better matched degrees because they are travelling further.

## 5.2 Degree subject studied

The subject the student chooses to study at university may be a factor in mismatch since, in the UK, students apply to specific university-subject combinations rather than universities. By conditioning on subject studied, we are exploring whether the SES and gender gaps are driven by low SES (female) students studying subjects with lower associated earnings than high SES (male) students, or whether they are attending lower quality institutions regardless of the subjects they study.

We see from Figure 8 that the inclusion of subject fixed effects does not substantially impact the SES gap parameter (e.g. for high attaining students it drops by around 1 percentile), meaning that little of the SES inequalities in match are driven by the subjects that people study at university. The implication is that even when they have similar qualifications, and are studying similar degree subjects, low SES students study at lower earning institutions. The same conclusion can be drawn using the achievement based measure of match (Appendix Figure A6), i.e. that even when they have similar prior achievement, and are studying similar degree subjects, low SES students study at less academically selective universities.

In contrast, Figure 9 shows that subject studied is *the only factor* that reduces the gender gap in match. For low-attainers, the gender gap reduces from 7.5 to 1 percentile, when estimated within subject grouping. This suggests that low-attaining women attend degrees in lower earning subjects compared to their male counterparts. For high-attaining students conditioning on subject studied reduces the gender gap to 2 percentiles, in line with our findings from Table 3. There are two points to note from these results:



First, subject choice is an important driver of the gender gap in earnings-based match. In contrast, we find that subject of study has no impact on the (admittedly small) gender gap in academic match (Appendix Figure A7). This implies that women attend degrees that are as equally academically selective as men, but which command lower earnings in the labour market. For example, highly qualified women may choose to study English at a selective institution, while men may choose a degree with an equally high entry requirement, but with higher potential earnings such as a STEM degree (Belfield et al, 2018). This is in line with the STEM literature (Card and Payne, 2017) which finds significant gender gaps in STEM entry. Critically, we are showing that this not driven by differences in the grades or subjects of the qualifications taken by men and women prior to university entry.<sup>23</sup>

Second, conditional on major chosen, females are still attending institutions with lower earnings potential. We find small but significant gender gaps in earnings-based match, for both high and low attaining women, even conditional on subject studied.

### 5.3 School characteristics

The final potential drivers of mismatch that we explore are secondary school characteristics. In Figures 8 and 9 we included an indicator for each secondary school. This school fixed effect will account for all school-level factors associated with the school, including information, peers, geography, and school sorting. The inclusion of school fixed effects greatly reduces the SES gap for both high- and low-attaining students, decreasing the SES gap by 73 and 79 percent respectively. This implies that much of the SES gap in match corresponds to these students attending different types of schools. Low and high SES students from the same secondary school tend to match to degrees in a more similar manner. However a significant SES gap still remains, with high-attaining low SES students enrolling in degrees with lower earnings potential than high SES students, by around 2 percentiles.<sup>24</sup>

We attempt to unpack the school fixed effect, by considering six different measures that could potentially relate to how well students from a school would match to university degrees. This is illustrated in Figure 11 (and for achievement-based match in Figure A9). The first three measures relate to the academic achievement of the school, including school-level

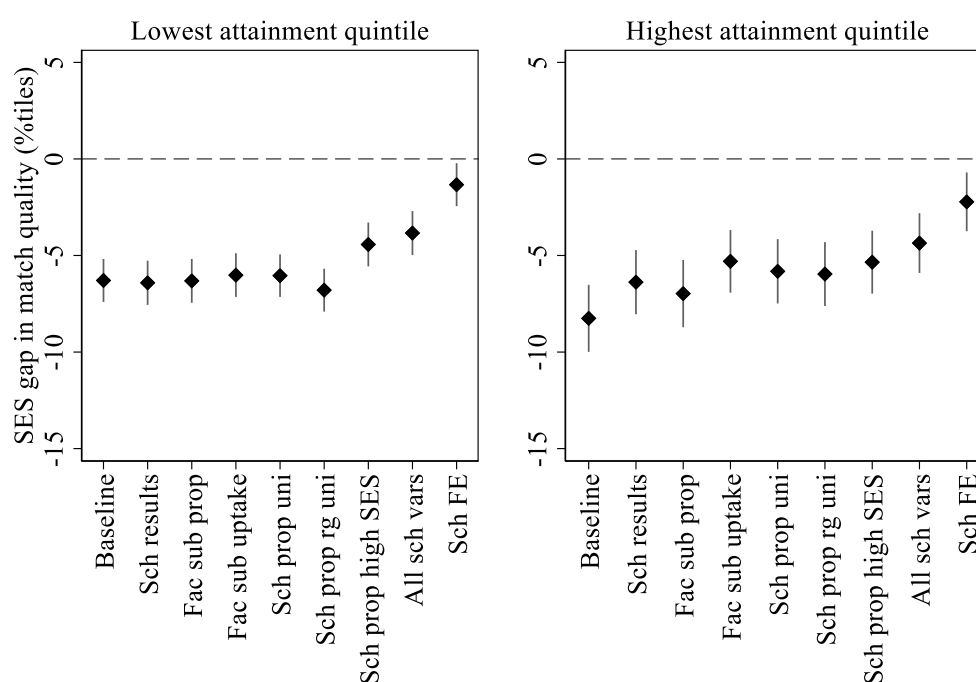
---

<sup>23</sup> The earnings gap is reduced throughout the student quality distribution once subject studied is taken into account. Figure A8 shows that a non-parametric plot by gender of students' national quality rank within major studied, and rankings of degrees within major (equivalent to Figure 3, but using the ranking within subject studied for degree) show only small gaps.

<sup>24</sup> Note from Figure 9 that school factors have no impact on the gender gap in match, which is expected as males and females are equally represented in most schools.

achievement, and the provision and uptake of qualifications in subject areas that are ‘preferred’ by universities. For high attaining students, school-level achievement explains around  $\frac{1}{4}$  of the SES gap in match, while the uptake of ‘preferred’ subjects accounts for an additional percentage point. Therefore students who attend schools where a higher proportion of the student body take more academically challenging qualifications are better matched to courses at university. The next three measures capture how much information about universities students at the school may be exposed to; the proportion of the school attending university, the proportion of the school attending high-status university, and the SES mix of the school attended (defined as the proportion of the school from the top SES quintile). While the first two do very little to explain the SES gap in match, the SES mix of the school attended does explain part of the SES gap for low attainers (around 2 percentage points), indicating that low attaining low SES students in schools with a greater proportion of high SES students are better matched than those with lower proportions of high SES students. Combining all six observable school characteristics accounts for around half of the SES gap for both low and high attainers, but cannot fully explain the total ‘effect’ of schools from the fixed effects model. This leaves scope for other unobservable features of schools, such as the information, advice and guidance offered, and the encouragement of teachers in these settings.

**Figure 11: SES gaps in earnings-based match, conditional on school factors**



## 6. Conclusions

Given the proven complementarities between student and course quality, the type of degree that students enrol in is an important factor in how much society benefits from higher education. Therefore from a social planner's point of view, ideally all students would be well matched to their degree. We document large inequalities in student-to-degree (university-subject) match using detailed administrative data from schools, universities and tax records, on some 140,000 students. We create two measures of match, one based on the academic achievement of students, and a new measure of match, characterising university degrees by the median earnings of graduating students.

We find a significant proportion of students are mismatched to the degree they attend. While a direct comparison with other studies is not possible, our results imply that there may be less mismatch in the UK than the US. This may be attributable to the UK's relatively generous financial system (students are eligible for maintenance loans, and fees are fully covered with income-contingent loans<sup>25</sup>), and the fact that there is almost no price variation between degrees, meaning poorer students cannot make a price-quality trade-off. The UK's centralised applications and admissions system, UCAS, which allows students to easily apply to up to 5 university degrees for a very small fee may also be a factor in helping UK students to match well to their degrees.

Yet despite these important features, we still find significant SES gaps in match. Low SES students more likely to undermatch and less likely to overmatch on academic-based match. This finding has been documented in previous papers in this area (Dillon and Smith, 2017; Smith, Pender and Howell, 2013). However our earnings-based measure of match shows that not only do disadvantaged students attend less academically selective degrees but they also enrol in lower earning degrees across the achievement distribution. This novel finding has important societal implications: if low SES students are attending degrees with lower returns, this will impact their future earnings, and undermine the potential for higher education to have a positive impact on social mobility. We find a key role for secondary school attended in accounting for our SES disparities in match, with the inclusion of school effects eliminating  $\frac{3}{4}$  of the gap. The academic rigour of schools attended explains part of this story for high-attaining pupils, while the social mix explains part of this story for low-attaining pupils. But around half of the school fixed effects are still unexplained, pointing to the role of other unobserved factors

---

<sup>25</sup> Barr et al (2019) considers the UK system more favourable to those in place in the US

associated with secondary school such as peers, parental sorting, and information provided by the school as likely key drivers for improving student-to-degree match.

Our earnings-match measure also highlights important gender gaps in match. In particular we find that women tend to choose degrees that are as academically selective as men, but with lower associated earnings. For both high-attaining and low-attaining women, subject choice plays a key role. But for high-attaining women, small gaps remain after controlling for subject studied: even where they enrol in a similar field as men, they still appear to study at institutions with lower average graduate earnings. This finding has implications for the gender pay gap, suggesting that higher education plays an important role in this much studied issue.

Recent studies have investigated the importance of providing information to low SES students to improve match (McNally, 2016; Dynarski et al, 2018). Our results highlight that it may also be beneficial to target women in a similar way, providing information on potential earnings associated with both institution and field of study. However, as with most studies of mismatch, we have no information on the preferences of students. Women may be well-informed on the earnings potential of subjects, but simply prefer not to study male dominated majors for rational reasons. This could be due to societal norms, or because women prefer to avoid being in male dominated industries. Or they may simply believe that the high average earnings associated with such subjects will not apply to them due to within-major gender gaps. Similarly, it may be the case that low SES students prefer to attend less academically challenging institutions even when their achievement levels suggest they are academically prepared.<sup>26</sup> Regardless, providing information, advice and guidance, in a targeted way that tries to break down existing barriers in terms of both understanding and perceptions, can only result in more informed choices.

---

<sup>26</sup> Sanders et al. (2018) found that using current students from elite university as ambassadors to dispel prospective low-income student's misconceptions increased applications and attendance to selective institutions.

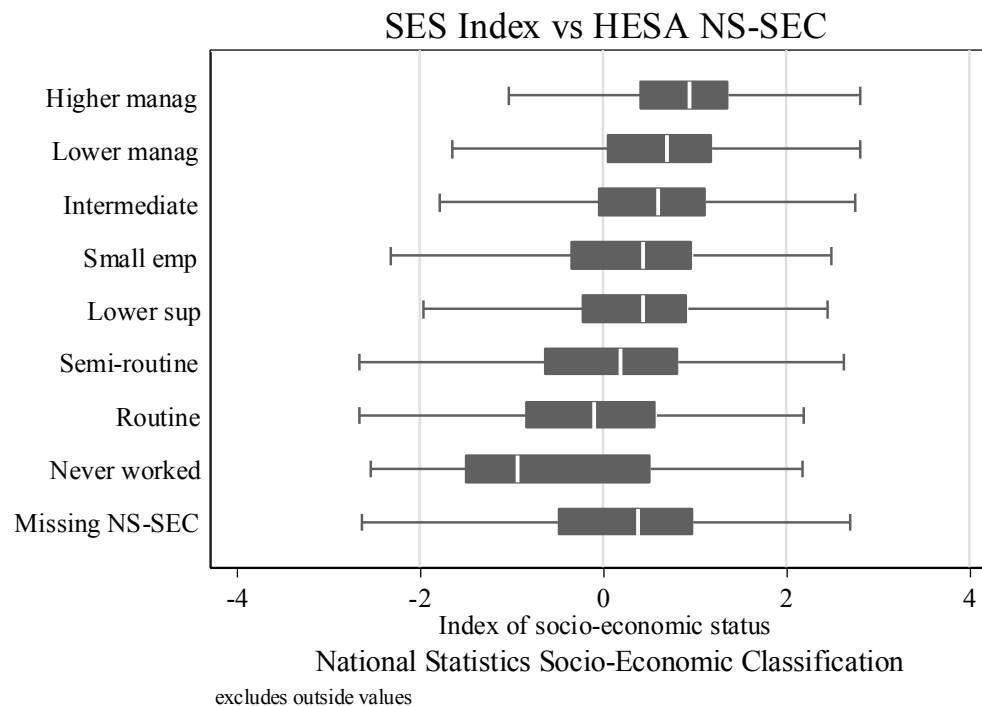
## References

- Avery, C. (2013). *Evaluation of the college possible program: Results from a randomized controlled trial* (No. w19562). National Bureau of Economic Research.
- Barr, N., Chapman, B., Dearden, L., & Dynarski, S. (2017). Getting student financing right in the US: lessons from Australia and England. *Centre for Global Higher Education*.
- Barr, N., Chapman, B., Dearden, L., & Dynarski, S. (2019). The US college loans system: Lessons from Australia and England. *Economics of Education Review*, 71, 32-48.
- Belfield, C., Britton, J., Buscha, F., Dearden, L., Dickson, M., Van Der Erve, L., Sibieta, L., Vignoles, A., Walker, I. & Zhu, Y. (2018). *The relative labour market returns to different degrees*. Institute for Fiscal Studies.
- Black, S. E., Cortes, K. E., & Lincove, J. A. (2015). Academic undermatching of high-achieving minority students: Evidence from race-neutral and holistic admissions policies. *American Economic Review*, 105(5), 604-10.
- Bolton, P (2019). Student loan statistics. Briefing Paper Number 1079, House of Commons Library, London
- Card, D., & Payne, A. A. (2017). *High school choices and the gender gap in STEM* (No. w23769). National Bureau of Economic Research.
- Carneiro, P., & Heckman, J. J. (2002). The evidence on credit constraints in post-secondary schooling. *The Economic Journal*, 112(482), 705-734.
- Chowdry, H., Crawford, C., Dearden, L., Goodman, A., & Vignoles, A. (2013). Widening participation in higher education: analysis using linked administrative data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2), 431-457.
- Coe, R., Searle, J., Barmby, P., Jones, K., & Higgins, S. (2008). *Relative difficulty of examinations in different subjects*. CEM Centre, Durham University.
- Department of Education (DfE). (2010). Schools, pupils and their characteristics. January 2010.
- Dillon, E. W., & Smith, J. A. (2017). Determinants of the match between student ability and college quality. *Journal of Labor Economics*, 35(1), 45-66.
- Dillon, E. W., & Smith, J. A. (2018). *The consequences of academic match between students and colleges* (No. w25069). National Bureau of Economic Research.
- Dilnot, C. (2018). The relationship between A-level subject choice and league table score of university attended: the 'facilitating', the 'less suitable', and the counter-intuitive. *Oxford Review of Education*, 44(1), 118-137.
- Dynarski, S., Libassi, C. J., Micheltore, K., & Owen, S. (2018). *Closing the gap: The effect of a targeted, tuition-free promise on college choices of high-achieving, low-income students* (No. w25349). National Bureau of Economic Research.
- Firpo, S., Fortin, N. M., & Lemieux, T. (2009). Unconditional quantile regressions. *Econometrica*, 77(3), 953-973.
- Gibbons, S., & Vignoles, A. (2012). Geography, choice and participation in higher education in England. *Regional science and urban economics*, 42(1-2), 98-113.

- Hoxby, C. M., & Avery, C. (2012). *The missing "one-offs": The hidden supply of high-achieving, low income students* (No. w18586). National Bureau of Economic Research.
- Hoxby, C. M., & Turner, S. (2015). What high-achieving low-income students know about college. *American Economic Review*, 105(5), 514-17.
- Kelly, A. (1976). A study of the comparability of external examinations in different subjects. *Research in Education*, 16(1), 37-63.
- Light, A., & Strayer, W. (2000). Determinants of College Completion. *Journal of Human Resources*, 35(2), 299-332.
- Lochner, L.J., & Monge-Naranjo, A. (2011). The Nature of Credit Constraints and Human Capital, *American Economic Review*, 101(6), 2487-2529
- McGuigan, M., McNally, S., & Wyness, G. (2016). Student awareness of costs and benefits of educational decisions: Effects of an information campaign. *Journal of Human Capital*, 10(4), 482-519.
- McNally, S. (2016). How important is career information and advice? *IZA World of Labor*.
- Murphy, R., Scott-Clayton, J., & Wyness, G. (2019). The end of free college in England: Implications for enrolments, equity, and quality. *Economics of Education Review*, 71, 7-22.
- Ofqual (2017). *Native speakers in A level modern foreign languages*. Office of Qualifications and Examinations Regulation.
- Oreopoulos, P., & Dunn, R. (2013). Information and college access: Evidence from a randomized field experiment. *The Scandinavian Journal of Economics*, 115(1), 3-26.
- Sanders, M., Burgess, S., Chande, R., Dilnot, C., Kozman, E., & Macmillan, L. (2018). Role models, mentoring and university applications-evidence from a crossover randomised controlled trial in the United Kingdom. *Widening Participation and Lifelong Learning*, 20(4), 57-80.
- Smith, J., Pender, M., & Howell, J. (2013). The full extent of student-college academic undermatch. *Economics of Education Review*, 32, 247-261.

## Appendix

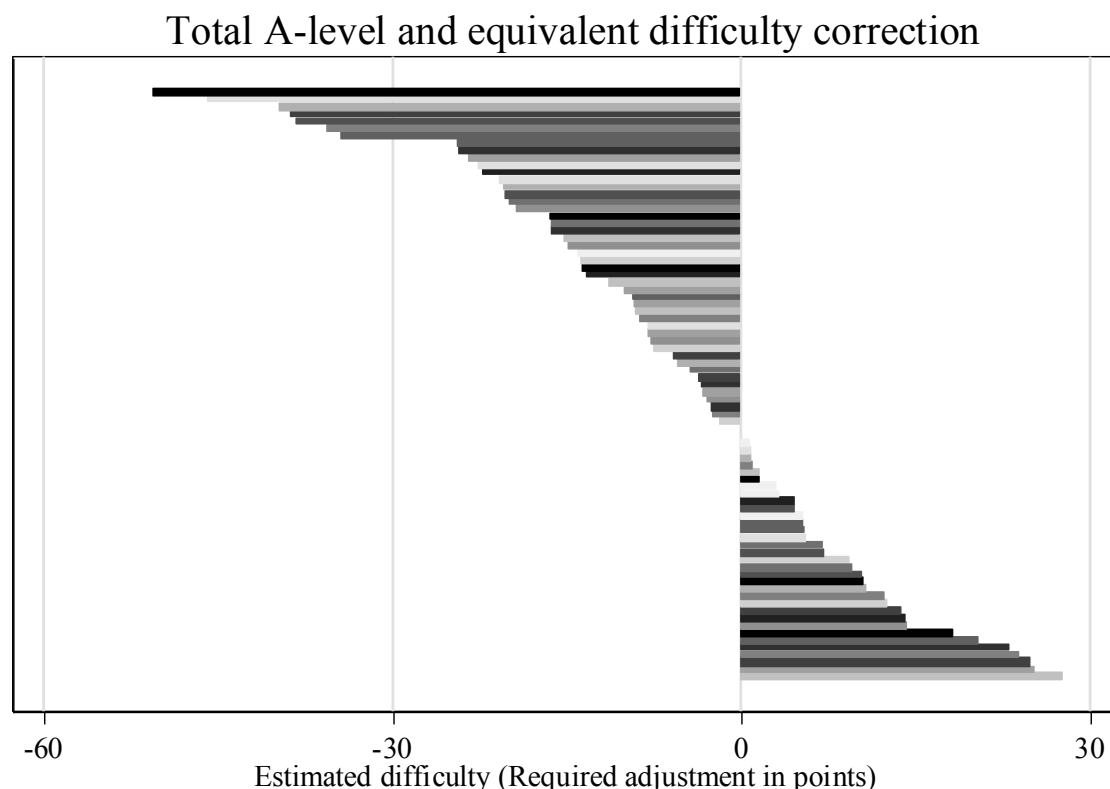
**Figure A1: Comparison of SES index with HESA NS-SEC**



Source: NPD-HESA. n=138,969.

Figure A1 compares our SES index with the ‘National Statistics-Socio-Economic Classification’ (NS-SEC) which is available for around 80% of university attendees in our sample. The NS-SEC measure available in the HESA data is a fairly noisy categorical indicator of SES, since it relies on a mapping from the parental occupation which each student enters into their university application form, and has a relatively high level of non-response. Still, it is reassuring that our continuous measure of SES places the categories of the NS-SEC in a plausible ranking.

**Figure A2: A-level subject difficulty correction**



Subject	Difficult y	Subject	Difficult y	Subject	Difficult y	Subject	Difficult y
Persian	-50.53	Dance	-15.23	PE	-3.39	IT	5.54
BTEC	-45.88	A&D: 3D design	-14.84	RE	-3.34	Ancient history	5.74
Bengali	-39.74	Drama	-14.09	Psychology (sci)	-2.90	Spanish	7.12
Urdu	-38.78	Fine Art	-13.82	Geography	-2.53	Accounting	7.25
Panjabi	-38.28	Art and Design	-13.64	Business studies and economics	-2.44	Mathematics (pure)	9.42
Turkish	-35.66	Portuguese	-13.26	English literature	-1.85	Greek	9.67
Polish	-34.43	Performing	-11.43	Psychology (soc)	-0.09	Computer studies	10.43
Film	-24.43	Dutch	-10.04	Home Economics	-0.02	Logic/philosophy	10.65
Communication	-24.31	D&T: Food	-9.27	Mathematics (statistics)	0.75	Mathematics	10.85
Russian	-23.44	Business	-9.20	Government and Politics	0.92	French	12.43
Modern Greek	-22.60	World Development	-9.05	Other classical languages	0.93	Music	12.72
A&D: Photography	-22.21	D&T: Production	-8.68	Archaeology	1.08	German	13.83
A&D: Graphics	-20.78	Electronics	-7.99	Modern Hebrew	1.64	Science (environmental)	14.24
A&D: Critical and contextual studies	-20.43	D&T: Systems	-7.97	Italian	1.67	Latin	14.33
Media, film, and TV	-20.30	English	-7.74	Gujarati	3.21	Science	18.39
A&D: Textiles	-19.89	Chinese	-7.51	Music technology	3.40	Biology	20.41
Arabic	-19.35	Law	-5.85	Classical civilisation	4.67	Mathematics (further)	23.12
Vocational double award	-16.45	English language	-5.50	History	4.71	Chemistry	23.94
Sociology	-16.27	Geology	-4.35	History of art	5.42	Physics	25.05
Japanese	-16.26	Vocational A-level	-3.70	Economics	5.51	Biology (human)	25.32
						Additional mathematics	27.74

Source: NPD-HESA. n=138,969. Units are QCA (Qualifications and Curriculum Authority) points, where one A-level grade is 30 points.

Our approach to accounting for varying difficulty in age 18 examinations across different subjects follows Kelly (1976) and Coe et al. (2008). We calculate difficulty scores for all subjects based on the full set of examination results for the population of students taking age 18 examinations in 2008. To calculate the difficulty of each subject, we subtract each

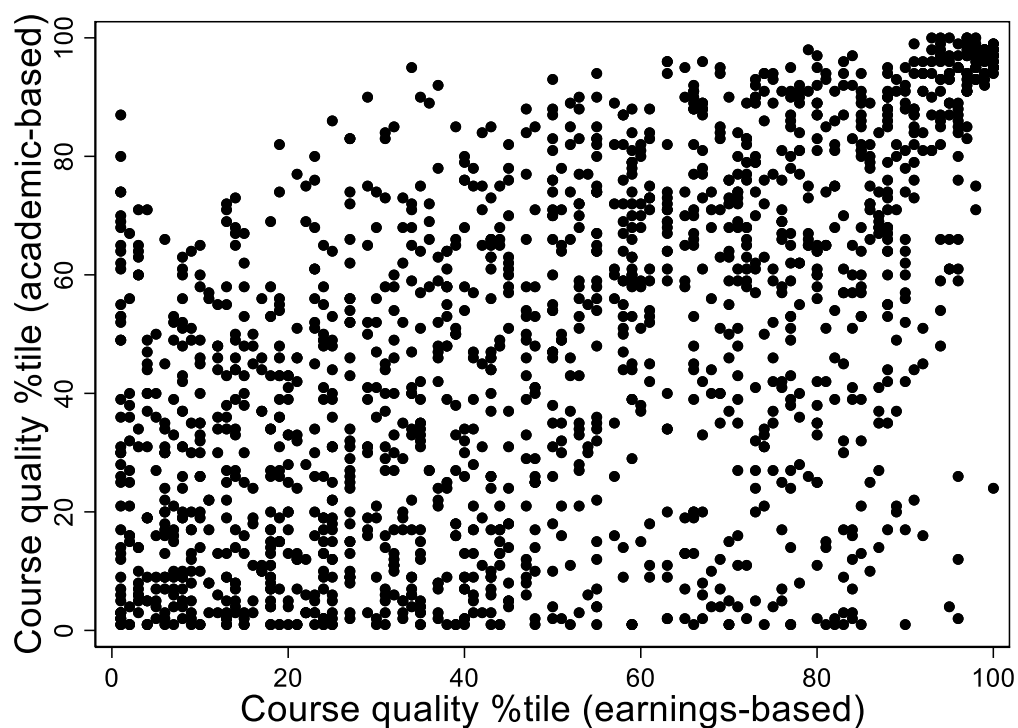


participating student's score in that subject from her average score across all other subjects, and sum the result across all participating students. This gives us a difficulty correction factor for each subject. As Kelly (1976) and Coe et al. (2008) note, these initial difficulty correction factors are likely to underestimate the variation in difficulty across subjects, since students who take 'hard' subjects tend to combine them with other 'hard' subjects, and those who take 'easy' subjects tend to combine them with other 'easy' subjects. We therefore 'correct' each student's score using the initial difficulty correction factors and repeat the process. We do this ten times. With each repetition, the difficulty correction get smaller, and after ten times, the required adjustments have effectively shrunk to zero.

Figure A2 and the accompanying table show the total difficulty correction factor applied to each subject. The units are "QCA points", and 30 points represents one A-level grade. The 'easiest' subject, Persian, with a total difficulty correction of -51, is therefore 2.6 grades easier than the most difficult subject, Additional Mathematics, which has a total difficulty correction of 28. Intuitively, this means that students who take Persian tend to score higher in that subject than they do in others, while those who take Additional Mathematics tend to score lower.

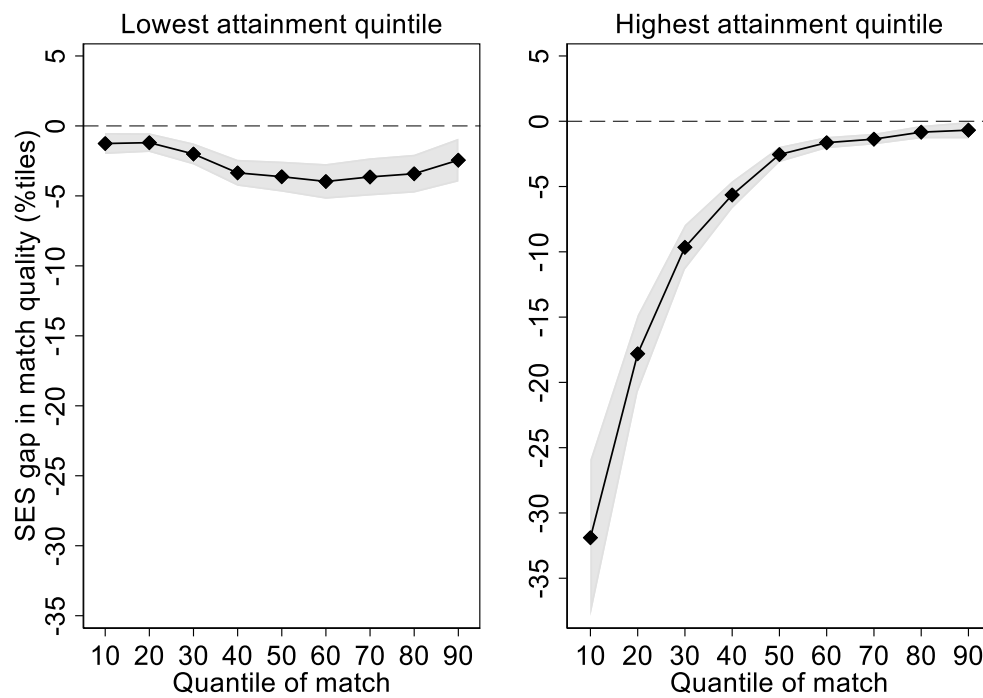
NB: In the case of Persian and other minority languages, it may be that many students who take the subject already have some understanding of the language, which decreases the perceived difficulty of the subject using our method (see Ofqual, 2017). Only a small numbers of students take these minority language A-levels – none of A-level Persian, Bengali, Urdu, Panjabi, Turkish, Polish, or Russian, is taken by more than 100 students in our estimation sample.

**Figure A3: Academic- and earnings-based measures of match**



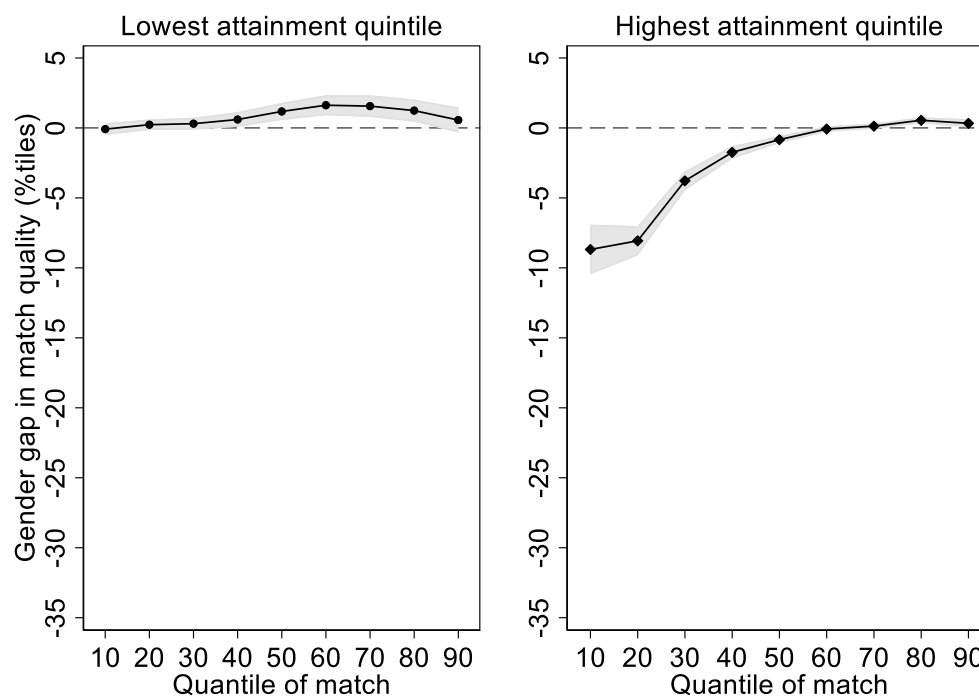
Source: NPD-HESA. n=1,722. Notes: Each point represents a university degree, plotted against our two quality measures: Median graduate earnings percentile (x-axis); Median student entry qualification percentile (y-axis).

**Figure A4: SES gaps in severity of academic-based match**



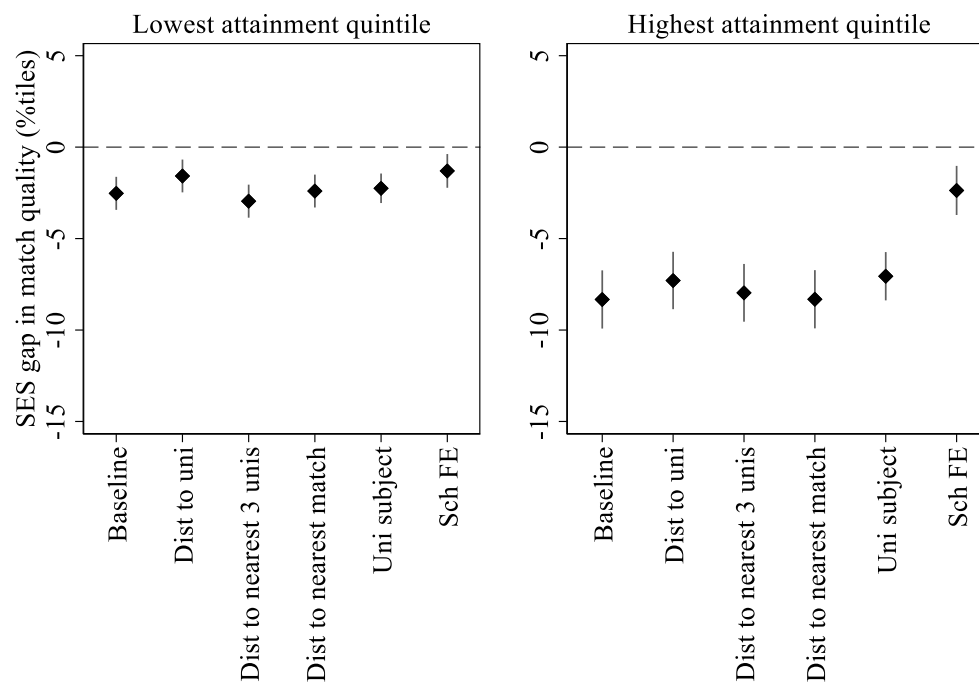
Source: NPD-HESA. n=138,969. Notes: Each point represents the SES match gap between groups 1 and 5 from specification 2, estimated for each decile of the match distribution. Controls are dummies for ethnicity, English as an Additional Language, Special Educational Needs, and gap year before college, and cubics in age 11 and age 16 exam results. We present the 95 percent confidence intervals, with standard errors clustered at the secondary school level.

**Figure A5: Gender gaps in severity of academic-based match**



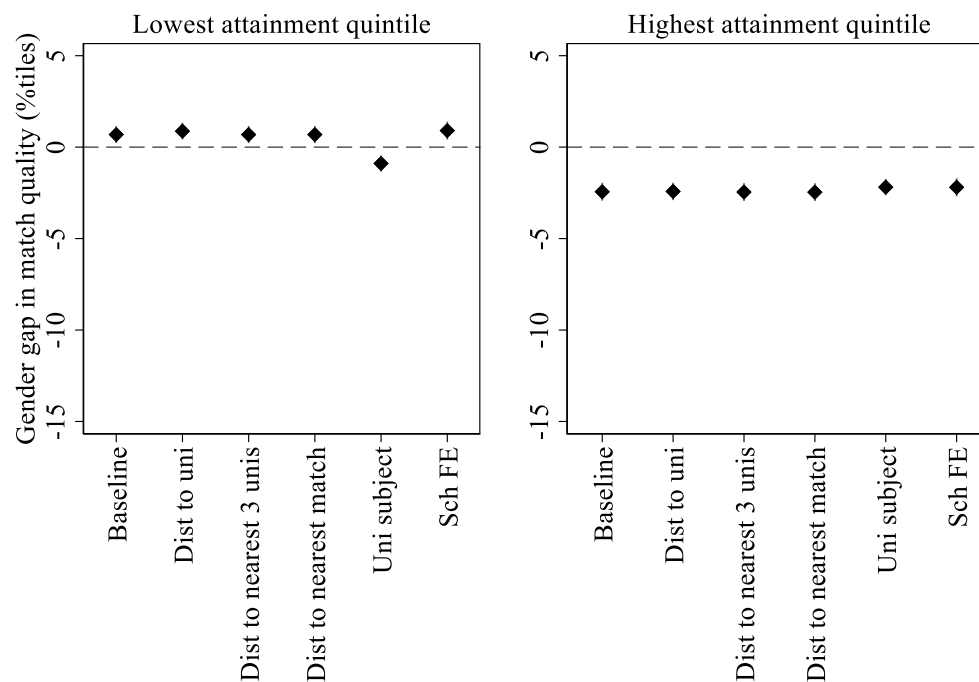
Source: NPD-HESA. n=138,969. Notes: Each point represents the gender match gap between groups 1 and 5 from specification 2, estimated for each decile of the match distribution. Controls are dummies for ethnicity, English as an Additional Language, Special Educational Needs, and gap year before college, and cubics in age 11 and age 16 exam results. We present the 95 percent confidence intervals, with standard errors clustered at the secondary school level.

**Figure A6: SES gaps in academic-based match, conditional on geography, subject, and schools**



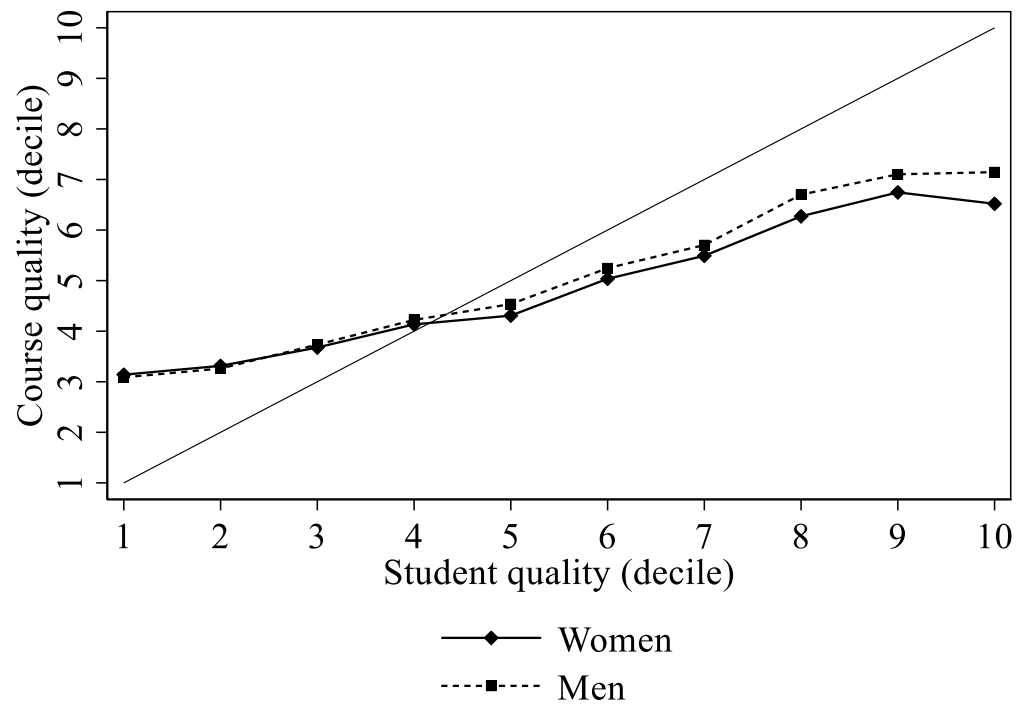
Source: NPD-HESA. n=138,969. Notes: Each point represents the SES match gap between groups 1 and 5 from specification 3, estimated for the top and bottom quintiles of the achievement distribution. The baseline controls are dummies for ethnicity, English as an Additional Language, Special Educational Needs, and gap year before college, and cubics in age 11 and age 16 exam results. We present the 95 percent confidence intervals, with standard errors clustered at the secondary school level.

**Figure A7: Gender gaps in academic-based match, conditional on geography, subject, and schools**

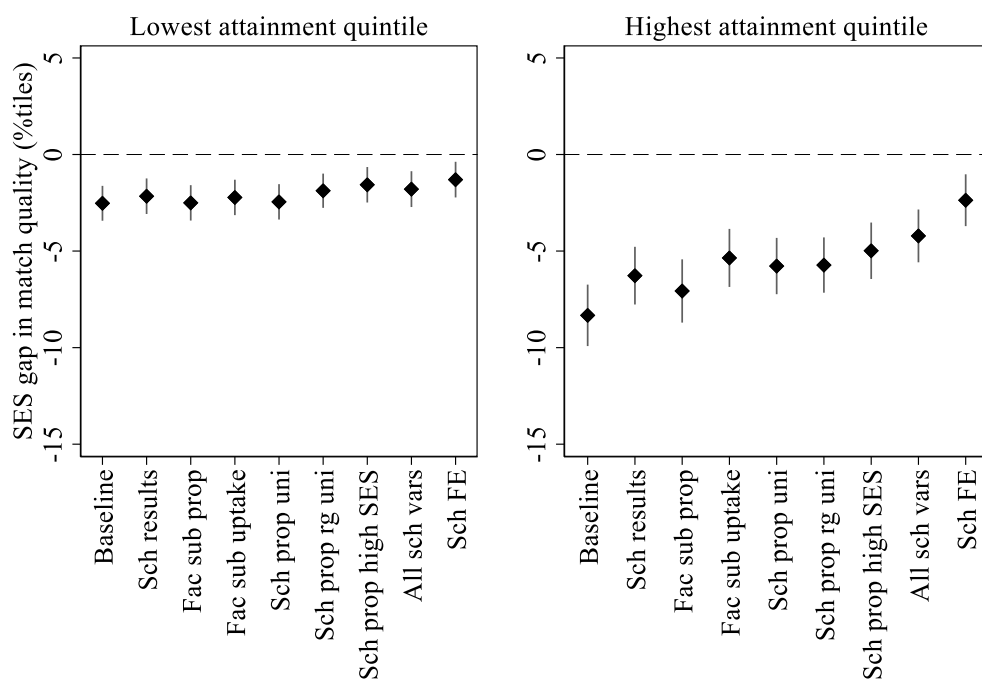


Source: NPD-HESA. n=138,969. Notes: Each point represents the gender match gap between groups 1 and 5 from specification 3, estimated for the top and bottom quintiles of the achievement distribution. The baseline controls are dummies for ethnicity, English as an Additional Language, Special Educational Needs, and gap year before college, and cubics in age 11 and age 16 exam results. We present the 95 percent confidence intervals, with standard errors clustered at the secondary school level.

**Figure A8: Gender earnings-based match by student achievement, within university subject studied**



**Figure A9: SES gaps in academic-based match, conditional on school characteristics**



Source: NPD-HESA. n=138,969. Notes: Each point represents the gender match gap between groups 1 and 5 from specification 3, estimated for the top and bottom quintiles of the achievement distribution. The baseline controls are dummies for ethnicity, English as an Additional Language, Special Educational Needs, and gap year before college, and cubics in age 11 and age 16 exam results. We present the 95 percent confidence intervals, with standard errors clustered at the secondary school level.



**Table A1: SES and Gender Conditional Match Gaps across Additional Alternative Specifications**

<b>Panel A: Academic-based match</b>												
	Baseline		No diff adjustment		KS2-based		Degree size weights		631 subject level		Combined hors	
Achievement quintile	1 <sup>st</sup>	5 <sup>th</sup>	1 <sup>st</sup>	5 <sup>th</sup>	1 <sup>st</sup>	5 <sup>th</sup>	1 <sup>st</sup>	5 <sup>th</sup>	1 <sup>st</sup>	5 <sup>th</sup>	1 <sup>st</sup>	5 <sup>th</sup>
SES quintile												
1	-2.53 (0.46)***	-8.33 (0.81)***	-1.75 (0.43)***	-8.88 (0.86)***	-1.43 (0.70)*	-3.26 (0.93)***	-2.58 (0.47)***	-7.89 (0.77)***	-2.05 (0.41)***	-7.27 (0.78)***	-2.49 (0.46)***	-7.76 (0.81)***
2	-2.42 (0.39)***	-4.47 (0.47)***	-1.58 (0.37)***	-5.05 (0.51)***	-2.04 (0.58)***	-1.75 (0.61)**	-2.48 (0.39)***	-4.21 (0.45)***	-2.28 (0.35)***	-3.81 (0.45)***	-2.16 (0.38)***	-4.40 (0.47)***
3	-1.69 (0.34)***	-3.29 (0.34)***	-0.99 (0.34)**	-3.55 (0.37)***	-2.60 (0.54)***	-2.59 (0.46)***	-1.75 (0.34)***	-3.10 (0.32)***	-1.35 (0.33)***	-2.94 (0.32)***	-1.57 (0.34)***	-3.12 (0.34)***
4	-0.62 (0.33)	-1.83 (0.27)***	-0.24 (0.33)	-2.13 (0.30)***	-0.89 (0.50)	-1.70 (0.43)***	-0.66 (0.33)*	-1.69 (0.26)***	-0.54 (0.32)	-1.64 (0.26)***	-0.50 (0.32)	-1.79 (0.28)***
Women	0.69 (0.24)**	-2.44 (0.25)***	1.48 (0.22)***	-3.32 (0.27)***	0.81 (0.38)*	-3.83 (0.36)***	0.72 (0.24)**	-2.27 (0.24)***	0.98 (0.23)***	-2.52 (0.24)***	0.89 (0.24)***	-2.27 (0.24)***
Constant	14.35 (0.33)***	-17.88 (0.57)***	8.64 (0.32)***	-15.43 (0.66)***	-7.35 (0.47)***	10.12 (0.35)***	17.19 (0.33)***	-17.06 (0.54)***	12.54 (0.31)***	-15.92 (0.54)***	13.61 (0.33)***	-17.78 (0.56)***
Clusters	2135	2005	2135	2005	2121	1983	2135	2005	2135	2005	2135	2005
n	27794	27786	27794	27786	26554	26580	27794	27786	27794	27786	27794	27786
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
<b>Panel B: Earnings-based match</b>												
Achievement quintile	1 <sup>st</sup>	5 <sup>th</sup>	1 <sup>st</sup>	5 <sup>th</sup>	1 <sup>st</sup>	5 <sup>th</sup>	1 <sup>st</sup>	5 <sup>th</sup>				
SES quintile												
1	-6.29 (0.57)***	-8.25 (0.88)***	-5.76 (0.57)***	-8.38 (0.96)***	-6.98 (0.75)***	-0.74 (1.10)	-6.14 (0.56)***	-8.14 (0.87)***				
2	-3.07 (0.48)***	-4.45 (0.53)***	-2.71 (0.49)***	-4.60 (0.57)***	-3.61 (0.67)***	0.09 (0.70)	-3.01 (0.47)***	-4.39 (0.52)***				
3	-1.72 (0.44)***	-3.89 (0.44)***	-1.43 (0.45)**	-3.92 (0.46)***	-2.46 (0.61)***	-1.88 (0.55)***	-1.69 (0.43)***	-3.84 (0.43)***				
4	-1.28 (0.44)**	-2.27 (0.36)***	-1.11 (0.45)*	-2.37 (0.39)***	-1.23 (0.60)*	-1.20 (0.51)*	-1.27 (0.43)**	-2.26 (0.36)***				
Women	-7.48 (0.32)***	-8.07 (0.32)***	-8.30 (0.33)***	-10.08 (0.34)***	-2.38 (0.42)***	-6.95 (0.43)***	-7.34 (0.31)***	-8.02 (0.32)***				
Constant	28.15 (0.47)***	-21.26 (0.56)***	27.81 (0.48)***	-13.08 (0.63)***	-1.81 (0.56)**	5.61 (0.40)***	28.17 (0.46)***	-21.77 (0.55)***				
Clusters	2135	2005	2135	2005	2121	1983	2135	2005				
n	27794	27786	27794	27786	26554	26580	27794	27786				
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes				

Source: NPD-HESA. n=138,969. All specifications control for ethnicity, English as an Additional Language, Special Educational Needs, and gap year before college, and cubics in age 11 and age 16 exam results. These prior achievement controls are omitted for the KS2-based regressions.

Appendix Table A1 presents a number of alternative specifications to supplement further the robustness checks discussed in Section 4. As discussed, we may be concerned that different students choose to study different A level subjects, which are not of equal value. To attempt to reduce the effect of this on our estimated SES and gender gaps, we adjust our total points by a subject difficulty rating. Our first additional robustness specification in columns 3 and 4 presents the SES and gender gradients without this adjustment, to show that it is making little difference to our estimates.

A further alternative method of dealing with the potential endogeneity of A-level subject choice is to alternative measures of student achievement to rank students and degrees (for academic-based match). We discuss the use of compulsory exams at age 16 in Section 4. Here we extend this even further to rank students based on their qualifications at age 11 (Key Stage 2). At age 11 all students in England take the same three exams in English, maths and science, which completely removes the issue of student choice. We sum the scores across subjects for each student and then calculate their national percentile rank. The degree ranks are in turn calculated on the basis of the median students using these measures. The results are similar for this very early ranking of student achievement, with a slight reduction in the SES gradient for high-attainers.

When calculating the percentile rank of the degree we weight degrees by the number of students in our administrative data. However, in some cases this will not include all students, as our data does not contain students that went to a private secondary school or are international students. Columns 7 and 8 therefore recalculate the degree percentile ranks using the actual number of students on the degree. This is not used for our main measure because we do not have data on the qualifications of these students. Therefore, for consistency we weight and rank degrees according to our population data. This makes very little difference to our estimates.

Throughout our analysis, we analyse match based on degree-level measures constructed from 23 broad subject levels across every institution. The reason for using 23 broad subject levels is that our earnings-based measure of ‘quality’ is only available at this level and we choose to keep the match measures consistent in this way. However for our academic-based measure, we have access to four digit JACS (Joint Academic Coding System) degree codes, which separately classify around 1,300 different university subjects. For example, we can separately identify those who are studying ‘Economics’ from those who are studying ‘Applied Economics’, and those who are studying ‘History by period’ from those who study ‘History by topic’. In Columns 9 and 10 we use our more detailed university\*subject groupings for our

academic-based measure of match to show that our results are robust to using the disaggregated subject categories.

Finally, columns 11 and 12 consider an alternative way to specific students on combined honours degrees. About 10% of our sample are doing these types of degrees, where each subject studied falls into more than one group of our 23 subject classification. In our baseline results we assign these students their highest weighted subject, or if weighting is equal they are given the first subject listed. In these results, we assign them according to their highest 2 weighted subjects, so there are 117 different categories including single and combined honours. Note it is not possible to carry out this test for our earnings-based measure because we do not observe later earnings for these combined honours degrees. Our results are again very similar using this alternative specification.

 @cepeo\_ucl

[ucl.ac.uk/cepeo](https://ucl.ac.uk/cepeo)