Is improving access to university enough? Socio economic gaps in the earnings of English graduates*

Jack Britton
Institute for Fiscal Studies, London

Lorraine Dearden
Institute for Fiscal Studies, London and Institute of Education, University College London

Neil Shephard
Department for Economics and Department of Statistics, Harvard University

Anna Vignoles
Department of Education, University of Cambridge

July 14, 2017

Abstract

Much research and policy attention has been on socio economic gaps in participation at university, but little attention has been paid to gaps in earnings. This paper addresses this shortfall using tax and student loan administrative data to investigate the earnings of English graduates up to their mid thirties by socio economic background. We find that graduates from higher income families (from the top fifth of the income distribution of those enrolled in university) have average earnings which are 20% higher than those from lower income families. Once we condition on institution and subject choices, this premium roughly halves, to around 10%. The premium grows with age and is larger for men, in particular for men at the most selective universities. We follow Chetty et al. (2017) and estimate English mobility scorecards by university and subject, highlighting the good performance of medicine, economics, law, business, engineering, technology, math, computer science and architecture courses as well as the prominent London-based universities.

Keywords: Administrative data; Graduate earnings; University access; Social mobility.

JEL: I23, I24, I26, J62

*Many civil servants and policy makers have helped us gain access to the data which is the core of this paper. Although it is difficult to pick out a small group who helped most, we must thank in particular Daniele Bega, Dave Cartwright, Nick Hillman, Tim Leunig and David Willetts who were all crucial in making this project happen. We solely are responsible for any errors. We are particularly grateful to the Nuffield Foundation for their financial support. The Nuffield Foundation is an endowed charitable trust that aims to improve social well-being in the widest sense. It funds research and innovation in education and social policy and also works to build capacity in education, science and social science research. The Nuffield Foundation has funded this project, but the views expressed are those of the authors and not necessarily those of the Foundation.

1 Introduction

It is well known that access to English universities varies substantially by the level of parental income, with many fewer students from poorer backgrounds attending higher status universities (Chowdry et al. (2013)). However, little is known about how these differences translate into differences in earnings between graduates from poorer and richer backgrounds. Further, primarily due to data limitations, the question of whether differences in earnings persist conditional on university and subject choice, has remained largely unanswered. In this paper we are able to address these shortfalls by making use a unique administrative database that tracks the earnings of graduates into their mid thirties.

We use a dataset described in Britton et al. (2015) that consists of anonymised individual level taxable earnings data supplied by Her Majesty's Revenue and Customs (HMRC), linked to information on their higher education (university or college) from the English Student Loan Company (SLC). The two data sets are hard linked using National Insurance numbers[1] and we have access to a 10% sample. We study cohorts of students who entered higher education from 1999-2005, and focus on earnings between 2008/09 through 2013/14. This allows us to follow graduates through their most crucial career developing years and well into their thirties. We also use Higher Education Statistics Agency (HESA) data which we can match at the subject-institution (rather than individual) level. This includes the socio economic background and pre-HE academic achievement of the students studying the same subject in the same institution. With these data we can add further controls that capture differences in the demographics of students in a given university and subject.

A typical problem with administrative data is a limited set of background characteristics for individuals.[2] We also face these limitations, and do not directly observe the parental income for individuals in our sample. However, we are able to infer a simple binary measure of parental income based on our SLC data since this includes the amount each student borrowed in their first year of study. For English students starting university before 2006, the amount individuals were eligible to borrow for their living costs was directly linked to their parental income. We identify people as being from a high income household if they are borrowing exactly the maximum amount an individual from a high income household is eligible for. This consists of approximately 20% of borrowers.

Our dataset is novel in a UK context. Whilst other UK surveys, such as the Labour Force Survey

[1] This is the key individual identifier for all taxes, social security and student loans.

[2] The availability of linked administrative data has improved dramatically in the UK in recent years. Future datasets will allow the linkage of entire education histories of individuals to their earnings records. These are not yet available to researchers, however.

(LFS) and the Destinations of Leavers from Higher Education (DLHE) survey, have information on subject of study and institution, higher education institution information has only recently been collected by the LFS, limiting the sample sizes available to researchers. The DLHE does have information on graduates' earnings by subject and institution but only captures full time equivalent earnings 3 years after graduation. Our administrative data therefore offers major advantages in terms of scale, quality and duration. However, aside from limited background characteristics it has two main drawbacks: we only observe borrowers, which means we do not see whether individuals graduate,[3] and we do not observe people who attend university but do not borrow; this accounts for 10-15% of all students and may be particularly weighted towards those from high income households.

Despite these caveats, we find considerable differences in earnings between graduates from richer and poorer family backgrounds. These differences roughly halve once we condition on subject and institution choices but remain economically important at around 10% and are statistically significant. These socio-economic differences also persist right through the earnings distribution and are larger at the bottom and top of the earnings distribution, suggesting family wealth is particularly good at both protecting graduates against very poor outcomes and providing them with opportunities for very high earnings. The conditional differences grow with age and are larger for "Other" subjects than for Science, Technology, Engineering and Mathematics (STEM) or Law, Economics and Management (LEM). They are also particularly pronounced for men from the most selective universities. We note that our results are descriptive, as crucially we are unable to control for individual cognitive and non-cognitive skills that may be correlated with labour market outcomes, while we only observe a blunt measure of parental income.

However, on balance, we believe this simply results in an underestimate of potential differences between rich and poor students, suggesting our findings are important and highly policy relevant. While much of the focus has been on improving attainment of poor students and access to higher education, these findings suggest that the policy challenge does not end there. Instead, our findings have implications for both universities, in relation to the career guidance and professional development support they give their students, and for firms, in regards to their hiring policies.

The most closely related paper to our work is Chetty et al. (2017), which investigates a similar phenomenon in the US using administrative data. They find somewhat different results, concluding that university does indeed level the playing field: graduates from poor and rich backgrounds have similar earnings following their higher education in the US. The different findings for England imply that there are differences in the higher education systems across the two countries or in the operation of labour markets that demand further research.

[3] We note that dropout from English universities is low by international standards at less than 10%. For simplicity we refer to borrowers as graduates throughout.

We also follow Chetty et al. (2017) by investigating mobility scorecards for English institutions, and unlike Chetty et al. (2017) we also provide them by subject choice. We find that medicine and economics are particularly good at delivering poorer students into the top 20% of the graduate earnings distribution. However, it is not at all clear that all STEM subjects are broadly effective at delivering social mobility. On the other hand LEM type subjects are. More broadly, professional facing subjects (e.g. LEM, math, computer science, engineering, technology, business, architecture) seem to deliver routes to social mobility. At the other end of the scale, biological sciences, mass communication and creative arts subjects and architecture perform relatively poorly. For institutions, the high-profile London universities - namely the LSE, Imperial College, King's College and UCL - do very well by this index, while outside of London, Warwick and Manchester are two of the best performing universities from the set we have permission to name.

The paper is laid out as follows. In the next section we discuss our contribution to the existing literature. In Section 3 we describe our data and introduce our measure of parental income. In Section 4 we presents results from our modelling. In Section 5 we follow Chetty et al. (2017) and present social mobility scorecards by subject and institution. Section 6 concludes and provides a further discussion of the policy implications.

2 Existing literature

This work will contribute to an important literature that has suggested a major impact from higher education on individuals' earnings. We focus on the English graduate labour market (Blundell et al. (2005)) though our findings are also relevant to the large US literature which has looked at the heterogeneity in graduate earnings by subject and institution (see for example Dale and Krueger (2014) and the survey by Altonji et al. (2015)). For English graduates, recent work by Walker and Zhu (Walker and Zhu (2011)) estimated the private lifetime earnings return to be in the order of £168k for men and £252k for women. Previous work using this dataset in Britton et al. (2015) also confirmed the major labour market advantage for English graduates over non graduates. For example, Britton et al. (2015) found that non graduates are twice as likely to have no earnings as graduates ten years after leaving higher education (30% against 15% respectively for the cohort graduating in 1999 and observed in the labour market in 2011/12). Further, Britton et al. (2015) found that half of non graduate women had earnings below £8k a year at around age 30, while only a quarter of female graduates were earning less than this. Similar patterns were observed for males. For those with significant earnings (which are defined as above £8k a year) median earnings for male graduates ten years after graduation were £30k, while the equivalent figure for non graduates was £22k. For women with earnings over the £8k threshold, median earnings ten years after graduation were £27k for graduates and £18k for non graduates. In summary, there is a sizeable wage gap between graduates and non graduates in England.

Yet although higher education in England appears to be a good investment for many, as is the case for the US there is also a sizeable empirical literature that has shown substantial variation in graduate earnings that has increased over time (Blundell et al. (2005), Bratti et al. (2005), Chevalier (2011), Hussain et al. (2009), Sloane and O'Leary (2005), Smith and Naylor (2001), Walker and Zhu (2011)). A key question is therefore, given this increased diversity in graduates' earnings, whether it is students from poorer backgrounds who are achieving lower earnings compared to their more advantaged but similarly qualified counterparts.

Differences in earnings between graduates from poorer and richer family backgrounds may of course be attributable to differences in the institutions they attend and the subjects they study. Previous work has shown that graduate earnings vary considerably by subject of degree (Sloane and O'Leary (2005), Chevalier (2011), Walker and Zhu (2011)). Walker and Zhu (2013) built on their earlier work (Walker and Zhu (2011)) which used the LFS to investigate how lifetime graduate earnings vary by both subject and institution type (as measured by broad university groupings, for example the "Russell Group", and "Millionplus" since at that time the LFS did not contain information on specific universities attended). Since they did not have access to survey data that contains both degree subject and institution for graduates of different ages, nor access to the administrative data we use here, they had to simulate the earnings profiles by splicing different survey data sets together (similar to work carried out by the Institute for Fiscal Studies on this issue, e.g. Chowdry et al. (2013)). Walker and Zhu (2013) suggested substantial differences in private returns by degree subject and insignificant differences in returns by institution type. However, they acknowledged that with the data they had available they were unable to test the robustness of these findings. Britton et al. (2016) also found considerable variation in earnings by both subject and institution, though much of this difference is attributable to different prior achievement levels of the students taking different degree options. Since prior achievement levels are lower, on average, for poorer students, we would expect sorting by subject and institution to depress their earnings.

Even with similar subject field and institution choices, an individual's socio economic background may have an effect on their labour market outcomes after graduation. This might be because students from more advantaged backgrounds have higher levels of (non-cognitive) skills (see for example Blanden et al. (2007)), i.e. skills that are not measured by their highest education level, or by their degree subject or institution. Alternatively, advantaged graduates may earn more because they have greater levels of social capital and are able to use their networks to secure higher paid employment. The literature in the UK at least does suggest that graduates from more advantaged backgrounds, particularly privately educated students, achieve higher status occupations and there is some evidence that privately educated students earn a higher return to their degree (Bukodi and Goldthorpe (2011b), Bukodi and Goldthorpe (2011a), Macmillan et al. (2013), Crawford and Vignoles (2014)). For example, Crawford and Vignoles (2014) found that graduates who attended private secondary schools earn around 7% more per year, on average, than state school students 3.5 years after graduation, even when comparing otherwise similar graduates and allowing for differences in degree subject, university attended and degree classification. This is consistent with earlier work using data from the 1970s and 1980s by Dolton and Vignoles (2000) that found the earnings return for graduates varied according to whether the individual attended a private school or a state school. This research also found that the private school wage premium for graduates who left university in 1980 was 7% for males but there was no premium for females, conditional on subject of degree and institution. Similar results were found by Naylor (2002) for a cohort of 1993 graduates (3% wage premium) and Green et al. (2012) using the National Child Development Study 1958 cohort and the 1970 British Cohort Study. The latter found that the private school wage premium increased from 4% for the earlier cohort to 10% for the later one. By contrast, work on how graduates' earnings vary by parental income level or parental socio-economic status, rather than by whether they attended private school, is more limited and Bratti et al. (2005) using the British Cohort Study (BCS) found little evidence of variation in the return to a degree by social class.

Perhaps the closest study to ours is Chetty et al. (2017) which looks at this issue for the US using administrative tax data linked to data from the National Student Loan Data System for around 30 million individuals who were university students between 1999 and 2013. Their study has the advantage of granular information on both parent and child income (the former measured when the student was aged 15-19 and the latter when the student was 32-34). From this, they were able to construct intergenerational income correlations for graduates from different institutions. They found stark differences in the likelihood of poor students accessing elite institution. For instance, a student with parents in the top 1% of the income distribution is 77 times more likely to go to an Ivy League university than those with parents in the bottom fifth of the income distribution. However, they also concluded that students from poor and rich backgrounds did similarly well if they graduated from the same college. Universities appear to be levelling the income playing field in the US. Our study has key differences from Chetty et al. (2017). First, our measure of parental income is binary which is a clear limitation. However, unlike Chetty et al. (2017) we are able to control for subject of study at the individual level, which is important given the evidence on variation in earnings by subject. Third, the English higher education admission system differs markedly from that of the US. Admission is centralised and regulated, with the probability of entry into elite institutions closely correlated with students' prior achievement in national examinations taken at age 18 (A levels or equivalent). In England it should not be possible, on average, for wealthy but low achieving students to gain access to elite institutions (though of course richer students are likely to have prior achievement levels due to having had substantially more investment than their poorer counterparts). Tuition fees in England are also supported by government provided income contingent loans, presenting less of a barrier to access compared to US loan systems. For England, Chowdry et al. (2013) find that conditional on prior achievement, there is no socio-economic gap on entry into HE and a gap of just a few percentage points on entry into elite universities. These different institutional arrangements may mean that the selectivity into HE and elite institutions is somewhat different in the two countries which will impact on graduates' earnings, particularly as both Chetty et al. (2017) and our own study are limited by not having individual level measures of skill or IQ. Hence the analyses in both papers is necessarily descriptive. Nonetheless, it is an important policy question as to the extent to which UK institutions appear to be levelling the earnings playing field in a similar way to their US counterparts.

3 Data

This paper uses a linked administrative dataset, accessed at a secure government datalab, that matches Her Majesty's Revenue and Customs (HMRC) earnings records with Student Loan Company (SLC) data on individuals' borrowing, university and subject choice. More extensive detail on the dataset is provided in Britton et al. (2015). The linked dataset includes official earnings data for a randomly drawn 10% sub sample of all borrowers from the English part of the SLC, which means they had to be domiciled in England upon application to university. We have data on those who entered higher education between 1998-2008 but focus on the 1999-2005 entrants (henceforth, 'cohorts') because of the low uptake of loans in 1998 (driven by the slow transition into the income contingent loan system) and the availability of tuition fee loans after 2005, which increased variation in borrowing amounts making it harder to identify individuals from higher income households.

These data provide us with information on gender, year the student first went into higher education (cohort), institution attended,[4] field of study,[5] region on application to higher education and a detailed measure of income from employment (Pay As You Earn taxable income) and from

[4]Students in officially recognised higher education learning institutions are eligible for loans. These are defined by the government as either 'recognised' or 'listed', the former can award degrees, the latter can offer courses that lead to a degree from a recognised institution. In our data students studying at both types of institutions will be observed. For example, students achieving their degrees at some Further Education Colleges will be included.

[5]This is at a broad subject level (first digit of JACS code, or 'JACS subject area' - see https://www.hesa.ac.uk/support/documentation/jacs/jacs3-principal).

self employment (Self Assessment income). We focus on earnings data from the tax years 2008/09 through 2013/14.[6]. Earnings are expressed in October 2012 prices using the Consumer Price Index.

Our information on graduates has some limitations. We have no information on non-English domiciled students, even if they work in the UK, since they do not take out loans from the English part of the SLC. We also only have information on the last institution the student went to, which means we cannot identify students who swap institutions during their time in HE. Neither of these factors are likely to strongly bias our estimates. However, we also cannot identify English domiciled students who chose not to take out a student loan. During this period, approximately 85% of English domiciled borrowers had taken out student loans. There is limited empirical evidence to help us understand which individuals do not take out loans, but one might anticipate that students who do not take out loans are likely to be more socio economically advantaged, attend higher status institutions and are more likely to go on to be higher earners. Any estimate of a socio-economic gap in earnings may therefore be an underestimate if the students from the wealthiest households are excluded from our sample.

The sample sizes for our cohorts of interest are given in Table 1, which also shows the gender split. These samples reflect 10% of English borrowers at UK Higher Education Providers. These sample sizes align with overall numbers from the Higher Education Statistics Agency (HESA) for the same period. There are more women, reflecting the higher participation rates of women in the UK (rather than different borrowing behaviour). Note that we use up to six years of earnings data for each individual throughout the majority of this paper.

Cohort	All	Men	Women
1999	22,621	10,590	12,031
2000	23,506	10,853	12,653
2001	23,924	11,025	12,899
2002	23,891	11,060	12,831
2003	23,972	11,024	12,948
2004	23,577	10,767	12,810
2005	25,103	11,439	13,664

Table 1: Number of graduates (10% sample of loan database), by cohort and gender. Cohort denotes the first year the individual received a loan from the SLC.

The administrative data described above is linked to data from HESA. Whilst we cannot link data at the individual level, we are able to do it at the institution and subject level. This provides a quantitative profile of the characteristics of students in each institution, subject and institution/subject combination. These data enable us to control more effectively for the characteristics

[6]The definition of earnings we use is detailed in Britton et al. (2015). We use earnings from labour, meaning employment income, profits from partnerships and profits from self-employment are included. We exclude trust income, profits on share transactions, profits from land and property, income from foreign employment, savings, UK dividends, pension income, life policy gains, "other" income, bank and building society interest.

of students attending different institutions and taking different subjects. This is important if we are trying to identify the residual correlation between socio-economic background and subsequent earnings after allowing for the fact that poorer students take different degree options. These data also allow us to control for the Government region in which the student's institution is located, which is important since wages vary by region and we do not have data on the graduates' current location (current region is in any case endogenous since graduates with degrees that are more highly valued in the labour market may be better able to secure high paying jobs in high paying regions). Since a high proportion of graduates remain near their university when they enter the labour market, controlling for region of institution goes some way to account for this issue. We use HESA data from 2002/03. The key characteristics for which we can control are: mean tariff score based on the UCAS 'tariff score' at the subject/institution level,[7] ethnic composition, gender composition and measures of students' socio-economic status. The latter include parental occupation, the percentage of students living at home whilst studying, the percentage of students who attended an English state school and the 'Participation of Local Areas' (POLAR) classification (neighbourhood level participation in higher education by age 19).

3.1 Creating our measure of parental income

Our focus is on how graduates' earnings vary by socio-economic background of the student. Unfortunately the data do not include a direct measure of parental income. However, the database does include the amount the graduate borrowed in their first year of borrowing. We are able to use this to make an inference about the parental income of each individual because the maximum amount the UK Government is willing to loan a student depends their parents' income, with individuals from lower income households able to borrow more than their more well-off peers. For the 1999-2005 cohorts that we investigate, there was a monotonic relationship between how much individuals could borrow and their parental income.[8]

There is a lot of noise in the observed amount individuals borrow. However, for each of the 1999-2005 cohorts we observe clear spikes at points in the distribution that we are able to exploit. To explain this, we provide an illustrative density plot in Figure 1. This shows the distribution of the amount individuals borrow in their first year of study, where x is the maximum an individual from a higher income household can borrow. The plot is normalised so x is set to 0 to allow for the fact that the maximum amount changes each year and differs for individuals inside and outside

[7]The tariff score is a single quantitative summary of the prior performance of students at Advanced level or equivalent.

[8]For subsequent cohorts this rule breaks down, as maintenance grants displaced some loans resulting in a non-monotonic relationship between parental income and loans. The introduction of tuition fee loans from 2006 adds an additional layer of complexity as many poorer students received grants to cover their tuition fees while richer students borrowed to cover their fee loans.

London.

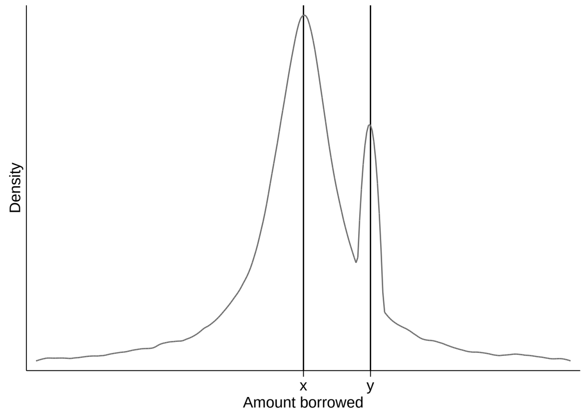


Figure 1: Illustrative density plot of amount borrowed. x represents the high income maximum, while y represents the low income maximum. Amounts and densities deliberately excluded for disclosure reasons.

The biggest spike in the distribution is at x, which captures around 20% of individuals. The second spike is at y, which is technically the most a low-income individual can borrow. Individuals borrowing more than this are low in number and typically have longer than standard course-lengths.

The exact loan amounts for each year, the minimum parental income threshold and the share of individuals at x (overall and by gender) are given in Table 2. The share of individuals borrowing exactly x in each year increases from around 15% in the 1999 cohort to around 25% in the 2004 cohort and averages around 20% across all cohorts. There are small gender differences, with a higher fraction of males borrowing x.

	Min Parental	Loan Amount		% borrowing = x		
Cohort	Income (£)	Non-London (£)	London (£)	Overall	Male	Female
1999	35,000	2,795	3,445	14.6	15.2	14.1
2000	36,000	2,795	3,445	18.9	20.2	17.8
2001	38,500	2,860	3,525	21.4	22.6	20.3
2002	40,000	2,930	3,610	21.8	23.2	20.5
2003	40,000	3,000	3,695	23.8	25.7	22.2
2004	40,950	3,070	3,790	24.8	26.1	23.6

Table 2: Loan amounts used to form the higher income household identifier and share individuals classified as higher income by gender. The parental income column gives the level of income (in nominal prices) above which individuals are eligible for a maximum of the loan amount given in columns 3 and 4 (depending on whether they are attending a London-based HEP).

Using this measure of borrowing we infer a blunt measure of parental income that we set equal to one (indicating high parental income) if the individual borrows exactly x in her first year, and zero otherwise (indicating low parental income). We acknowledge that this measure of parental income does not perfectly identify all student from higher income households, for a number of reasons. First, those from higher income households may borrow less than the maximum available.

Second, individuals from low income households may choose to only borrow the higher income maximum. Third, we are missing altogether those individuals from the wealthiest households who did not borrow at all. While this figure is around 10% of the overall student population, it is likely to represent considerable fractions of the student populations at some high-status institutions in particular. However, although this is an imperfect measure, we provide robust evidence below that it does indeed identify individuals from more wealthy households. Further, we suggest that all of these issues with the measure are likely to bias our impacts towards zero.

3.1.1 Validation of the parental income indicator

Here we investigate whether our simple indicator is indeed picking up high income individuals by showing how it relates to university access and voluntary repayments. First, we show the share of high income students in different types of institutions. We know that poorer students on average access less selective universities where the mean entry tariff score is lower. We divide all the universities in our database up into deciles,[9] in terms of the mean entry tariff scores of their students, based on HESA data. We then plot, in Figure 2, the percentage of students in each of those groups that comes from a high income background, by gender. It is clear that for both men and women, universities with higher entry criteria have much higher shares of individuals we define as being from a high income household. In the most selective universities, more than half of students come from the 20% of individuals we define as being higher income.

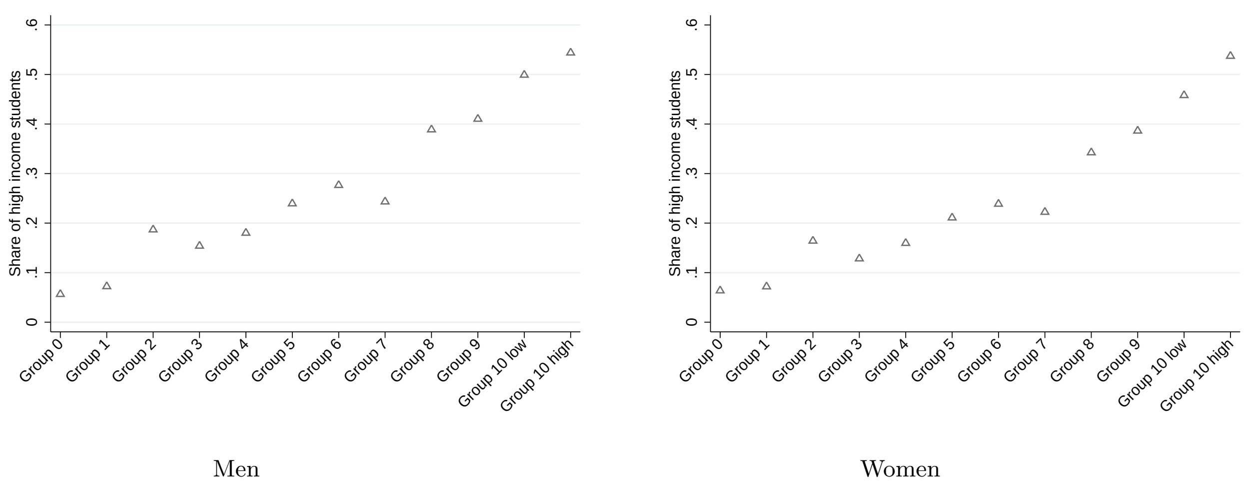


Figure 2: Share of individuals in each university group at the High Income HH borrowing amount. Includes the 1999-2005 cohorts. Shares incorporate borrowers only.

Second, we examine the voluntary repayments of students. These are repayments of student

[9]Note Group 1 is the bottom decile of university based on tariff entry, while Group 10 is the top. Group 10 is split into two groups so that Group 10 high is the top 5% of universities. Group 0 includes institutions where the entry tariff of students is not known.

loan amounts that are made over and above the legally required repayments that are determined by income. They are tracked by SLC and recorded in our database. A summary of the voluntary repayments is given by gender in Table 3. Around 9% of students make voluntary repayments at some point in our SLC dataset (specifically, between starting university and 2011)[10] and the mean annual repayment (conditional on making a repayment) amount is considerable. A marginally higher share of women make repayments than men, and women on average make more voluntary repayments, with 34% of those making any repayments making more than one, versus 29% for men. However average repayments are typically smaller for women than they are for men.

	Men			Women		
	Mean	S.D.	N	Mean	S.D.	N
Share ever making repays	0.08	-	*76,776*	0.09	-	*89,818*
Share making >1 repayment	0.29	-	*6,460*	0.34	-	*7,991*
Share making >2 repayments	0.13	-	*6,460*	0.16	-	*7,991*
Share making >3 repayments	0.06	-	*6,460*	0.08	-	*7,991*
	£			£		
Average repay overall	2,840	4,071	*9,544*	2,490	3,785	*12,632*
Average repay 4 years from starting	5,058	4,696	*419*	5,041	4,613	*608*
Average repay 5 years from starting	4,720	5,222	*813*	4,037	4,913	*1,182*
Average repay 6 years from starting	3,817	4,795	*1,015*	3,208	4,399	*1,485*
Average repay 7 years from starting	3,146	4,422	*1,249*	2,478	3,802	*1,722*
Average repay 8 years from starting	2,494	3,668	*1,224*	2,144	3,491	*1,684*
Average repay 9 years from starting	2,407	3,715	*1,283*	2,066	3,383	*1,740*
Average repay 10 years from starting	2,292	3,413	*1,406*	1,891	3,071	*1,712*
Average repay 11 years from starting	1,873	2,956	*1,042*	1,776	2,906	*1,245*
Average repay 12 years from starting	1,998	3,269	*716*	1,503	2,478	*831*
Average repay 13 years from starting	1,577	2,789	*377*	1,616	2,791	*423*

Table 3: Voluntary repayments summary statistics

In Table 4 we estimate the probability of individuals from higher income households making any voluntary repayments. We estimate a probit model with a dummy set equal to one if an individual makes any repayment in a given year. The results show that individuals we classify as being higher income are significantly more likely to make voluntary repayments, even once conditional on their current earnings. They are about one percentage point more likely to make voluntary repayments, on a baseline of 3.3%.

	Repayments [1]	Add Gender [2]	Add Earnings [3]	Add HESA [4]
High Income HH	.189***	.191***	.182***	.142***
	(.007)	(.007)	(.007)	(.007)
$P(repays > 0 \mid$ Low income HH)	.033	.033	.033	.033
$P(repays > 0 \mid$ High income HH)	.048	.048	.047	.044
N	666,376	666,376	666,376	666,376

Table 4: Probit regression predicting ever making repayments. *** indicates significant at the 1% level; ** the 5% level. Controls for cohort, age and year are included in all columns.

[10]For clarity, the complete SLC dataset includes information on borrowers between 1998 and 2011, although we only use the 1999-2005 cohorts. We have earnings data until 2013/14, but that is HMRC rather than SLC data.

Table 5 further investigates voluntary repayments by investigating differences in the size of individual repayments. The table shows results from regressing the individual voluntary repayments made by students on demographic characteristics and the high income household indicator. Individual repayments from those from high income households are considerably larger than for those from lower income households. Again, this holds true when controls for gender and current earnings are added. Among those who make voluntary repayments, those from high income households make repayments that are around £1,000 larger on average. When HESA controls for subject-institution mix of students doing the same course are included, this estimate reduces to around £600, but remains statistically significant. This strongly favours the argument that individuals borrowing exactly x are indeed from more advantaged households than those who borrow different amounts.

	Repayments [1]	Add Gender [2]	Add Earnings [3]	Add HESA [4]
High Income HH	976.9***	958.5***	966.9***	614.5***
	(55.8)	(55.8)	(55.9)	(59.9)
Female		-384.6***	-395.0***	-334.2***
		(51.3)	(51.4)	(53.9)
Earnings			-0.004**	-0.008***
			(0.001)	(0.001)
Constant	2624.6***	2849.8***	2949.1***	2502.9***
	(72.5)	(78.4)	(85.5)	(720.3)
N	22,176	22,176	22,176	22,176
Adjusted R^2	0.067	0.069	0.070	0.095

Table 5: Size of total voluntary repayments (£), conditional on making them. *** indicates significant at the 1% level; ** the 5% level. Female is a dummy set equal to one for women. Controls for cohort, age and year are included in all columns.

The table also shows that conditional on making repayments, women make smaller repayments by around £330 on average, while the relation between voluntary repayments and current earnings is economically immaterial, despite being statistically significant (the earnings coefficient in column 4 suggests a £10,000 increase is associated with a reduction in voluntary repayments of just £8).[11]

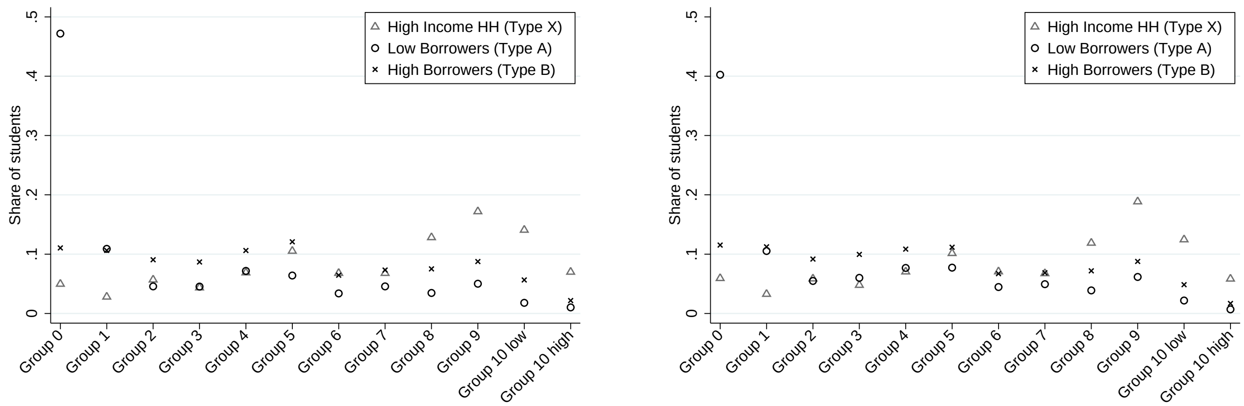
3.1.2 Treatment of those borrowing $< x$

We also investigate closely those who borrow less than x (as in Figure 1) to best determine how they should be treated. We repeat the above analysis, splitting out those who borrow below x (Type A) and above x (Type B) from those who borrow exactly x (Type X).

Figure 3 shows the distribution of university attendance for the three groups, split by gender. Note that this differs from Figure 2 by showing the density function for each of the three groups so that the total for each group sums to one. The most notable feature is the high share of Type A borrowers in Group 0. This is the group of institutions that do not have an average tariff score,

[11] Note that involuntary repayments are closely related to earnings, as repayments are income contingent.

and typically consists of smaller lower-status universities and Further Education colleges. Beyond that, it is clear that Type A individuals look much more like Type B individuals than they do Type X individuals. A very low share of Type A and B individuals attend the top 30% of universities, with a tiny fraction going to the top 5%. This contrasts with Type X individuals, of whom a high share go to top institutions.



Men

Women

Figure 3: Share of individuals attending different university groups from different parts of the borrowing distribution.

In Table A1 in the Appendix, we investigate voluntary repayments of Type A and B individuals relative to Type X. Both make much smaller voluntary repayments than Type X individuals, with Type A individuals making smaller voluntary repayments than Type B individuals. Of course, Type A individuals have lower debt, which makes them less likely to make large repayments. However, this is a big difference compared to Type X individuals, and is suggestive that they again are more like Type A individuals than Type X individuals. Based on this, we treat Type X individuals as our 'High Income Household' group and our Type A and B individuals as our 'Low Income Household' group. We investigate the robustness to this assumption in our subsequent analysis. We now move on to consider the raw earnings differences between individuals from the two groups.

3.2 Descriptive earnings differences

Figure 4 shows the earnings distribution for male and female graduates from high income households (grey triangles), graduates from low income households (black circles) and for non graduates (grey line), for the 1999 cohort in 2012/13. The non graduate sample comes from the HMRC databases (more information is given in Britton et al. (2015)). Points to the right of each figure show the mean for each group. The results are striking; graduates from higher income households earn more right across the distribution, from the 20th percentile upwards, for both females and males. Whilst graduates from both lower and higher income households earn more than non graduates, the gap between graduates from lower and higher income backgrounds is also sizeable, particularly at the very top of the distribution. Indeed, whilst around 20% of the graduate population come from high income households by our definition, of those in the top 1% of the earnings distribution, 45% (men) and 39% (women) come from high income households.

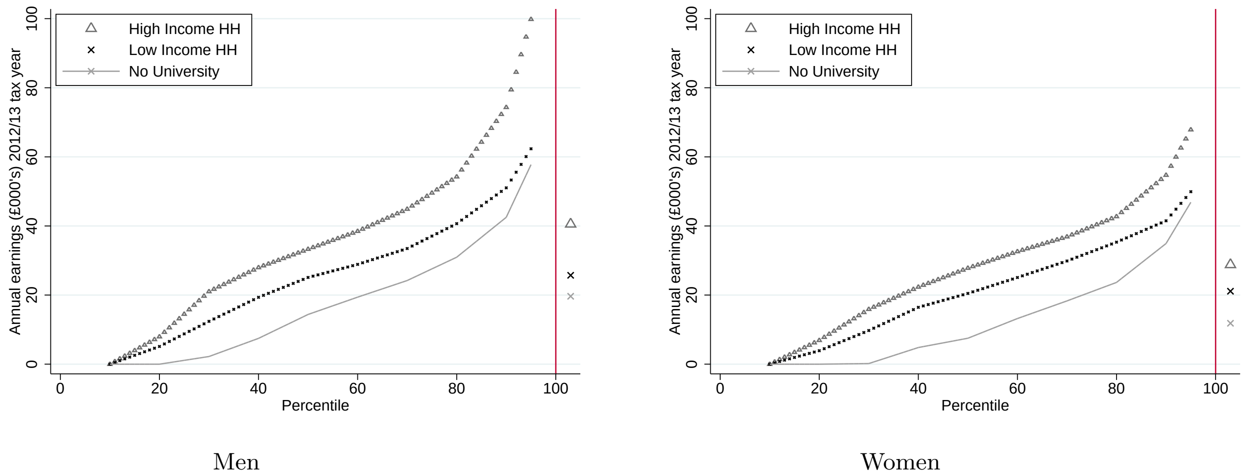
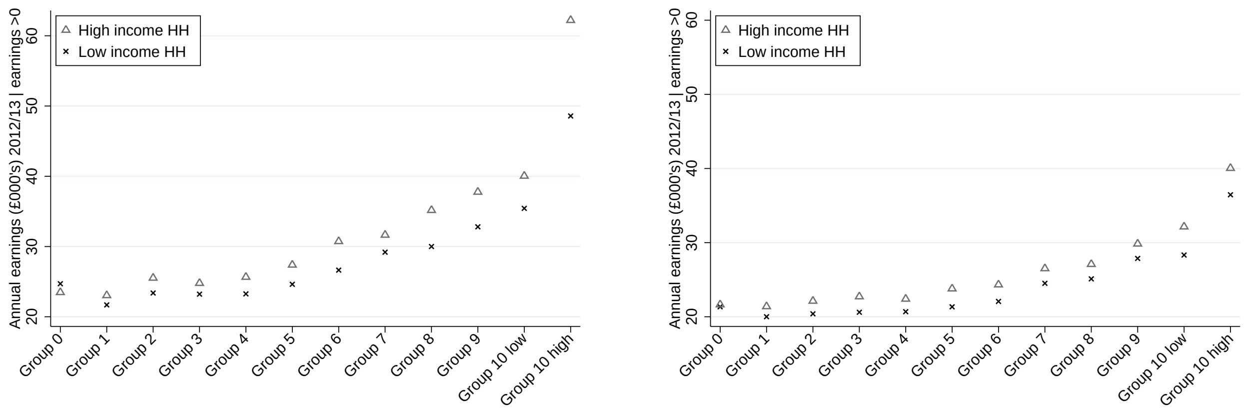


Figure 4: Earnings distribution for individuals in the 1999 cohort in 2012/13 for those from higher income households vs. individuals from lower income households, with the non-university distibution also included. Means given on the RHS of the plot.

As already discussed, students from different socio economic backgrounds take different degrees, with students from higher income households more likely to attend high status universities. It is possible that this sorting into universities could explain the raw earnings differences between those from high and low income households. Figure 5 takes the first step to address this by plotting average earnings (conditional on earnings being positive) for graduates by the university groups defined above, by gender. Even within these institution groups, the differences in average earnings between graduates from high and low income households are clear, suggesting that broadly speaking even when comparing graduates from the similar institutions, those from a higher income background go on to do better in the labour market. This appears to be particularly pronounced for men from the most selective universities.

Of course, these figures do not properly control for different degree choices between those from high and low income backgrounds. Individuals from higher income households might attend the more selective institutions within our coarse university grouping, or might make subject choices that lead to higher earnings. In the next section we try to address this more formally by investigating earnings differences conditional on subject and institution, as well as some other demographic characteristics.



Men

Women

Figure 5: Average earnings (given earnings > 0) for graduates from higher and lower income households by university tariff group.

4 Estimation

In Table 6 we estimate the following, conditional on individuals having positive earnings[12]

$$ln(y_{it}) = \alpha + \beta H_i + X_{it}'\gamma + \epsilon_{it} \tag{1}$$

where y_{it} is earnings of individual i at time t, H_i is an indicator for whether an individual is from a high income household and X_{it} is a vector of controls. We sequentially add additional controls into the vector X.

Columns 1 and 2 indicate the raw differential in earnings between students from higher income households, conditioning only on cohort and year. The differences in earnings are sizeable at around 21% for men and 16% for women. Controlling for subject of degree in columns 3 and 4 reduces these premia by 1-2 percentage points, suggesting that choice of subject explains very little of the differences in earnings. Adding variables which control for the different characteristics of students attending a particular degree course reduces the coefficients considerably (columns 5 and 6). This implies that the nature of the degree course, particularly the entry tariff score, explains more of the variation in earnings between high and low income students than does their choice of subject. In the final column we include university fixed effects. This does not make an appreciable impact on the coefficients, over and above controlling for the characteristics of the students attending a particular degree course. Overall the results indicate that even allowing for both institution and subject, students from higher income households earn around 10% more than students from lower income households. This suggests that higher education does not fully level the playing field in

[12] Alongside this approach, we also estimated a probit model to investigate differences in the share with zero earnings in a given year. We find negligible differences between the groups. See Appendix.

terms of graduates' earnings.

	Unconditional		Plus subject		Plus HESA		Plus HEI	
	Men	Women	Men	Women	Men	Women	Men	Women
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
High income HH	0.205***	0.157***	0.185***	0.150***	0.108***	0.0970***	0.113***	0.102***
	(0.004)	(0.003)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Cohort poly	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subject controls	No	No	Yes	Yes	Yes	Yes	Yes	Yes
HESA controls	No	No	No	No	Yes	Yes	Yes	Yes
HEI controls	No	No	No	No	No	No	Yes	Yes
Adjusted R^2	0.089	0.059	0.110	0.077	0.121	0.083	0.126	0.089
N	399,463	470,272	399,463	470,272	399,463	470,272	399,463	470,272

Table 6: Regression of log earnings on high income dummy and various controls. Includes the 1999-2005 cohorts and earnings data between 2008/09 and 2013/14. *** indicates significant at the 1 percent level. Standard errors are clustered at the individual level.

We assess the robustness of our findings to different definitions of high income in Table 7. Defining a student as being from a high income family in a number of different ways, we still obtain the same broad result which is that there remains a wage premium coming from a higher income household of approximately 10%, even conditioning on degree subject and institution. The alternative definitions of high income student are as follows, where x and y are defined in Figure 1:

- Baseline definition: amount borrowed $= x$;
- Definition 1: amount borrowed $\leq x$;
- Definition 2: amount borrowed $< y$;
- Definition 3: amount borrowed $= x$, but individuals with amount borrowed $< x$ excluded.

	Baseline		Definition 1		Definition 2		Definition 3	
	Men	Women	Men	Women	Men	Women	Men	Women
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
High income HH	0.113***	0.102***	0.099***	0.086***	0.089***	0.075***	0.126***	0.115***
	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)	(0.004)	(0.004)
Cohort poly	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subject controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
HESA controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
HEI controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.126	0.089	0.126	0.089	0.126	0.089	0.141	0.101
N	399,463	470,272	399,463	470,272	399,463	470,272	283,904	327,578

Table 7: Robustness of earnings regression to high income definition. Includes the 1999-2005 cohorts and earnings data between 2008/09 and 2013/14. *** indicates significant at the 1 percent level. Standard errors are clustered at the individual level.

Another robustness check is presented in Table 8 which compares OLS regression estimates to those obtained using a nearest-neighbour propensity score matching estimate. To deal with convergence issues in the HRMC datalab, we use the specification from columns 5 and 6 from Table 6 and match on the same set of variables as included in the OLS equation. Again the results are very similar. For men, the coefficient increases marginally by one percentage point, while for women it reduces slightly. Hence even with an alternative, arguably more flexible estimation approach, we find that the wage premium for students from rich backgrounds is around 10%.

	Baseline OLS		Matching	
	Men	Women	Men	Women
	[1]	[2]	[3]	[4]
High income HH	0.108***	0.0970***	.118***	.094***
	(0.004)	(0.004)	(.011)	(.009)
Cohort poly	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes
Subject controls	Yes	Yes	Yes	Yes
HESA controls	Yes	Yes	Yes	Yes
HEI controls	No	No	No	No
N	399,463	470,272	399,463	470,272

Table 8: Matching estimates of differences in log earnings. Includes the 1999-2005 cohorts and earnings data between 2008/09 and 2013/14. *** indicates significant at the 1 percent level. Standard errors are clustered at the individual level.

4.1 Heterogeneity

We are interested in potential heterogeneous effects, particularly across different subject areas. It may be that the advantage of high family income impacts upon some subject-occupation trajectories more than others. Table 9 shows the preferred specification but estimated separately for three different subject areas, namely LEM (Law, Economics and Management courses), STEM (Science, Technology, Engineering and Mathematics courses) and Other (the rest, typically humanities, languages and the arts). The wage premium from coming from a rich household is similar across all three subject areas except for women who take STEM subjects where interestingly the premium is somewhat lower at around 7%.

	LEM		Other		STEM	
	Men	Women	Men	Women	Men	Women
	[1]	[2]	[3]	[4]	[5]	[6]
High income HH	0.104***	0.101***	0.117***	0.111***	0.106***	0.0675***
	(0.009)	(0.008)	(0.006)	(0.005)	(0.006)	(0.008)
Cohort poly	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Subject controls	Yes	Yes	Yes	Yes	Yes	Yes
HESA controls	Yes	Yes	Yes	Yes	Yes	Yes
HEI controls	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.161	0.134	0.107	0.077	0.128	0.096
N	70,524	72,915	170,598	271,052	158,341	126,305

Table 9: Regression of log earnings on high income dummy and controls by subject group. Includes the 1999-2005 cohorts and earnings data between 2008/09 and 2013/14. *** indicates significant at the 1 percent level. Standard errors are clustered at the individual level.

Another aspect of heterogeneity we are able to explore is the magnitude of the wage premium from coming from a high income family for those who attend different institutions. Figure 6[13] shows the wage premium for the groups of institutions we discussed earlier, namely split into deciles by mean tariff entry score. It is striking that for males only, the wage premium for those from high income backgrounds is considerably larger if the student attended an institution in the top 5% of the institutional distribution, at around 25%. For women this effect is not evident.

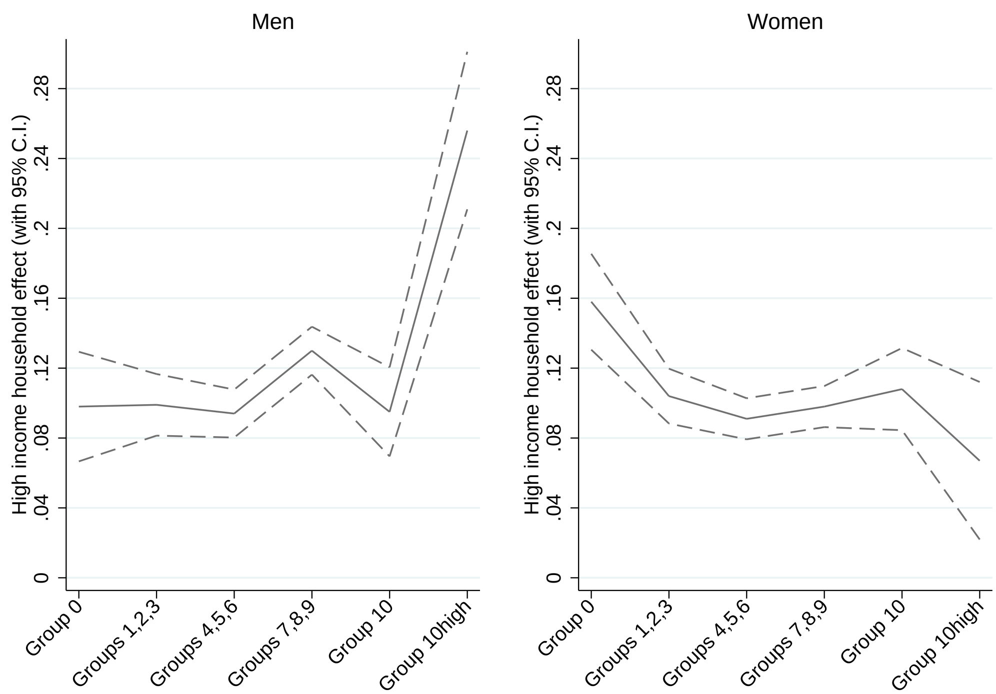


Figure 6: High income household effect split by university grouping. Includes the 1999-2005 cohorts and earnings data between 2008/09 and 2013/14. Dashed lines represent 95% confidence intervals. Corresponding tables given in the Appendix.

We also explore heterogeneity by age. Because we cannot disentangle age from cohort and year effects, we show cohort effects holding year fixed in Figure 7, and year effects holding cohort fixed in Figure 8. It is evident that the wage premium from coming from a high income background increases in both cases, suggesting that the impacts increase with age. Typically they appear to rise to around 14% for men and 12% for women by graduates' early thirties, starting at around half that in each case in graduates' mid twenties. There appear to be gender differences in how the effect changes with age; for women, there appears to be a dip in both figures at points that correspond to their early thirties, potentially due to family formation decisions. For men, this dip is not present, with the effects apparently continuing to rise.

[13]Full tables for these results are given in the Appendix.

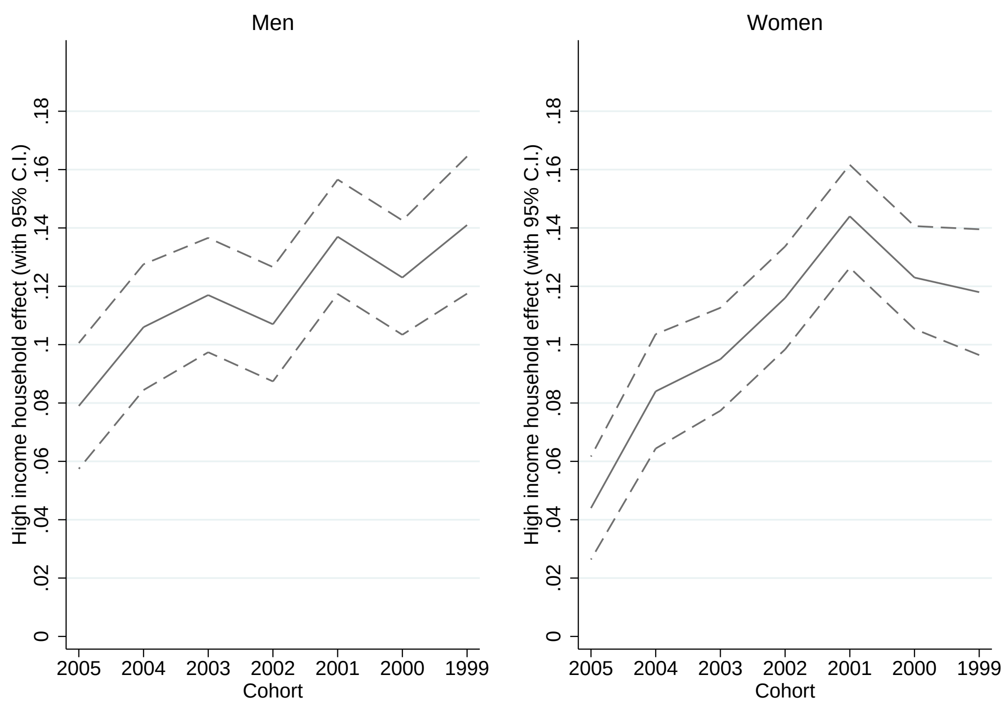


Figure 7: High income household effect split by cohort (inverted, to show increasing effect with age). Includes earnings data between 2008/09 and 2013/14. Dashed lines represent 95% confidence intervals. Corresponding tables given in the Appendix.

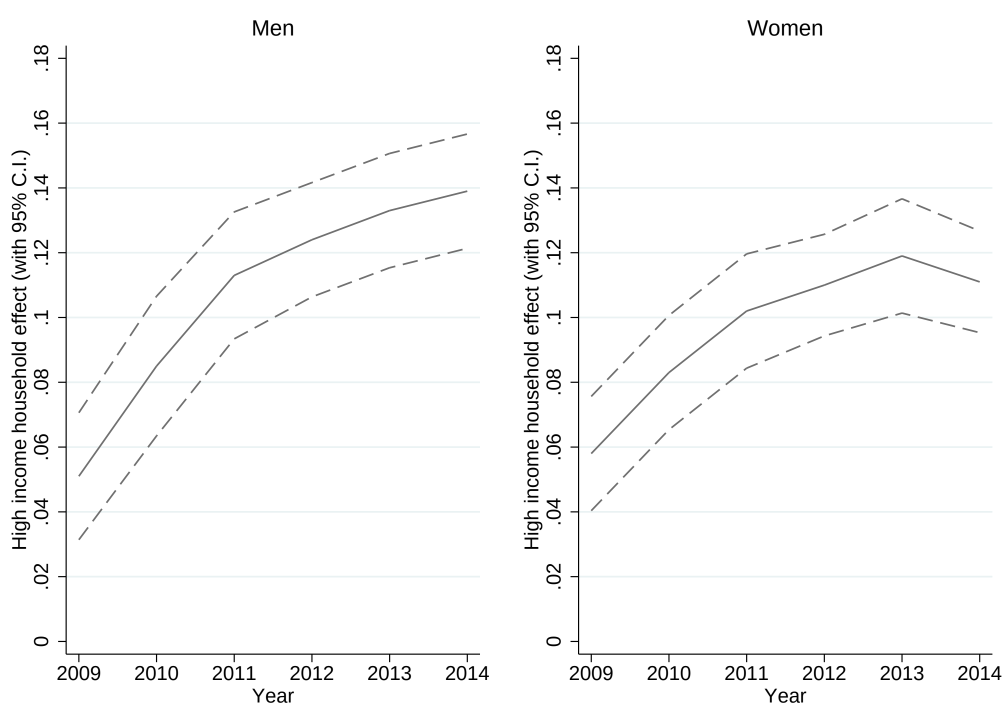


Figure 8: High income household effect split by year. Includes the 1999-2005 cohorts. Dashed lines represent 95% confidence intervals. Corresponding tables given in the Appendix.

Finally, in Table 10 we investigate the magnitude of the wage premium at different quantiles of the earnings distribution. Due to issues with convergence of the estimator, we use a restricted dataset that includes only the 1999 cohort in 2011/12 and 2012/13. We provide raw earnings and conditional estimates at the 20th, 50th and 90th percentiles of the distribution, separately by gender. What is striking is that although at the median the conditional wage premium for men and women is around 10%, this rises to 16-20% for men at the bottom (20th percentile) and the top (90th percentile) of the distribution. A similar, though less stark pattern is present for women. This implies that those from higher income households are both better protected against low earnings and more likely to achieve high earnings.

	Raw earnings (£000's)		Raw differences from low family income (£000's)		Conditional differences from low family income (£000's)	
	Men	Women	Men	Women	Men	Women
			20th Percentile			
High Income HH	20.3	14.0	7.7***	4.4***	3.0***	1.7***
	1.1	*.7*	*.9*	*.6*	*.7*	*.6*
Low Income HH	12.6	9.6				
	.5	*.3*				
% Wage Premium	61.1	45.8	61.1	45.8	16.1	13.4
			50th Percentile			
High Income HH	35.0	27.8	8.0***	5.3***	3.3***	2.1***
	1.0	*.7*	*.9*	*.6*	*.6*	*.6*
Low Income HH	27.0	22.5				
	.5	*.4*				
% Wage Premium	29.6	23.6	29.6	23.6	10.9	8.6
			90th Percentile			
High Income HH	84.0	54.9	30.8***	13.5***	10.7***	6.2***
	7.0	*2.1*	*6.8*	*2.0*	*1.8*	*1.3*
Low Income HH	53.2	41.4				
	1.5	*.8*				
% Wage Premium	57.9	32.6	57.9	32.6	19.6	14.3
Years	Yes	Yes	Yes	Yes	Yes	Yes
Cohort	Yes	Yes	Yes	Yes	Yes	Yes
Regions	No	No	No	No	Yes	Yes
Age	No	No	No	No	Yes	Yes
Subject	No	No	No	No	Yes	Yes
HEI	No	No	No	No	No	No
HESA	No	No	No	No	Yes	Yes
N	18,038	20,413	18,038	20,413	18,038	20,413

Table 10: Earnings differences for graduates from lower and higher income households at the 20th, 50th and 90th percentiles estimated from quantile regression models. Note that zero earnings are excluded from these regressions. High family income premium indicates the additional earnings for graduates from a higher income household. Low family income earnings indicates earnings of graduates from a lower income background. Percentage wage premium calculates the wage premium for those coming from a richer family background compared to the earnings of those from lower income households, assuming all controls are held constant across the two groups at their means. The first two columns of results show raw estimated earnings for high and low household income earnings. The next two show the difference in earnings from low household income. The final two columns show the conditional difference from low household income - i.e. the difference once controls for region, age, subject and student characteristics are included. All figures are in £000's. Uses 2011/12 and 2012/13 data and the 1999 cohort (estimates are given for 2012/13). Standard errors are clustered at HEP level. * indicates significantly different to the base (lower family income) at 10% level, ** 5% and *** 1%.

4.2 Summary

In summary, men from higher income households earn around 21% more than men from lower income households, while the equivalent figure for women is 16%. These estimates roughly halve to around 10% once controls for university, subject and other demographics are included. The differences appear to increase with age, doubling between individuals' mid twenties and their early thirties, before levelling off for women but continuing for to rise for men. This suggests previous work which has focussed on socio economic differences in early career outcomes (for example, Macmillan et al. (2013); Crawford et al. (2016)) may underestimate earnings gaps. Given our data limitations on the career paths of these individuals, this encourages further research. In particular it would be interesting to determine whether earnings differences by family background in graduates' early thirties are driven by initial placement into careers with faster earnings trajectories, or whether differences persist, even conditional on early career choices.

The socio economic gap in graduates' earnings is similar across broad subject groupings, with the exception of women doing STEM courses, for whom the earnings gap is considerably smaller. This latter result may be attributable to the types of occupations pursued by women in STEM, particularly those in medicine and the public sector where salaries are more regulated and hence where coming from a more advantaged family may make less difference to earnings. It could also be a selection effect, if women pursuing STEM are somewhat atypical and if family background makes less of a difference to these atypical women in their career prospects.

By university group, we find very large earnings differences amongst men from the highest status universities of around 25% holding actual institution and subject choice constant. This is a stark finding. It suggests that coming from a high income household and attending a top institution provides a disproportionate opportunity to achieve very high earnings for men and aligns with our quantile regression results.

The overall earnings differences we observe are large, particularly given the bluntness of our parental income measure. Indeed we believe this bluntness is likely to result in an underestimate of the true difference. This is because some lower income individuals will borrow the high income maximum, while some higher income individuals will borrow less than their full allocation. Further, some individuals will not borrow at all; and we would expect these to be especially from higher income households. All of these issues are likely to bias downward our estimates.

Some caveats apply. We cannot rule out the possibility that the earnings differences observed are due to individual characteristics that we are unable to control for. Although we do condition on subject-institution level attainment, we do not have any measures of individuals' prior cognitive or non-cognitive skills, that may be correlated with both their degree success and their future earnings. It is not necessarily clear which direction the bias would go, however. For example, individuals from lower income households may be more motivated and driven to achieve than their higher income peers doing the same courses. Further, high status degree courses in England generally have strict entry criteria for entry and these criteria are harder to achieve for students from poorer backgrounds (Chowdry et al. (2013)).[14] This means that poor students in such institutions may have been disproportionately able and motivated and this would bias their earnings upwards, suggesting we may be understating the socio economic gap in graduates' earnings. Equally, it may be that poor students lack other skills that are associated with high earnings (such as non cognitive skills) and hence we are overstating earnings gap.

5 Mobility scorecards

The results above suggest that, unlike in the US, universities do not appear to be levelling the playing field in terms of earnings. In this section, we follow Chetty et al. (2017) and estimate mobility scorecards to consider what these findings imply for social mobility. Specifically we investigate the extent to which different subject and university choices appear to help individuals from lower income backgrounds to become top earners, defined as having earnings in the top quintile of the earnings distribution. We split this analysis out by gender, although we consider the probability of getting to the top 20% of the overall earnings distribution, pooled across genders.[15]

Specifically, Chetty et al. (2017) define a mobility score card as follows:

$$P(\text{Child in Q5 and Parent in Q1}) = P(\text{Parent in Q1}) \text{ x } P(\text{Child in Q5} \mid \text{Parent in Q1})$$

Where Q5 is the top quintile of the income distribution and Q1 is the lowest quintile. As discussed, our data on parental income is not as rich as theirs. We therefore estimate the probability of a child making it to the top quintile of the earnings distribution given they are from a lower income household. Figures 9 and 10 follow Chetty et al. (2017) by plotting, for men and women respectively, $P(\text{Child in Q5} \mid \text{Lower income household})$ on $P(\text{Lower income household})$ for 21 subject groups we observe in our data (see the Appendix for more information on the subject groupings). For each subject, we plot the rank on the overall scorecard. This gives a sense of how good different subject groups are at delivering individuals that come from lower income households to the top of the graduate earnings distribution. Note carefully that our 'lower income household' individuals

[14] While in recent years there has been increased flexibility from some institutions to let poorer students in with slightly lower grades through contextualised admissions, this was not a prominent feature of the system for the cohorts we are investigating.

[15] Results looking at the gender-specific distributions are available on request from the authors.

make up 80% of our population of students.

Medicine and economics are the highest performing subject groups by this measure, not least because these are subjects in which a higher proportion of students (irrespective of their background) have very high earnings. Although medicine and economics are amongst the worst performing subjects in terms of the proportion of students enrolled from lower income backgrounds (65%), their delivery of students into the top 20% of the earnings distribution is very good. At least 40% of lower income students taking these subjects get into the top 20% of the overall earnings distribution. Other high mobility subject groups are maths and computing and engineering and technology. Miscellaneous law economics and management subjects also do relatively well.[16]

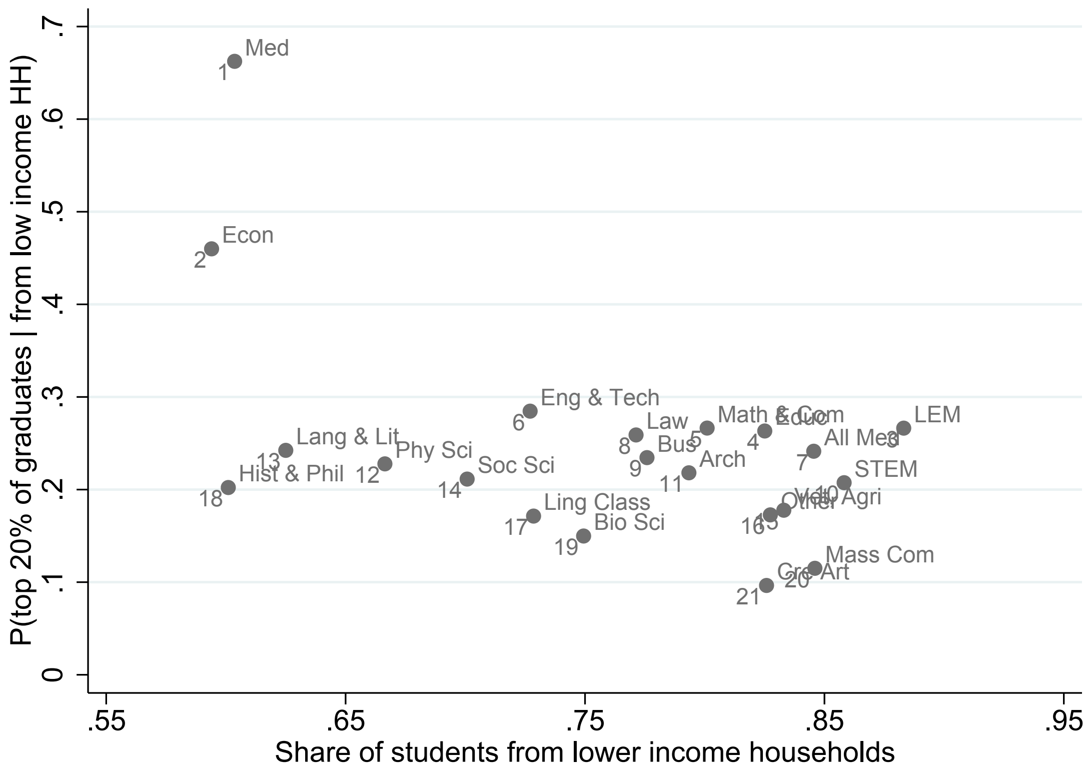


Figure 9: Subject mobility scorecard for men. Earnings rankings use the 2011/12-2013/14 tax years, treating individual observations as independent. The results are not very sensitive to this approach, however. The numbers represent within-gender subject ranking by the mobility index, which is the product of the x and y axes. See Appendix for full subject definitions. Numbers behind this figure are available in the Appendix.

On the other hand, we see languages and literature, history and philosophy, linguistics and classics and biological and physical sciences all have a relatively low share of lower income students enrolled and also have very poor delivery of those students into the top of the earnings distribution. Though creative arts does far better at enrolling students from lower income backgrounds, it is the

[16] This is a broad subject group and includes students whose subject group at their given institution was too small for us to get their detailed subject grouping.

worst subject in terms of enabling students from lower income households to reach the top of the earnings distribution. This latter result is because more generally students taking creative arts are less likely to achieve very high earnings, rather than being attributable to some failure within this subject for poorer students to thrive. Nonetheless, from a social mobility perspective, it is clear that some subjects are more likely than others to provide a pathway for poorer students to achieve very high earnings.

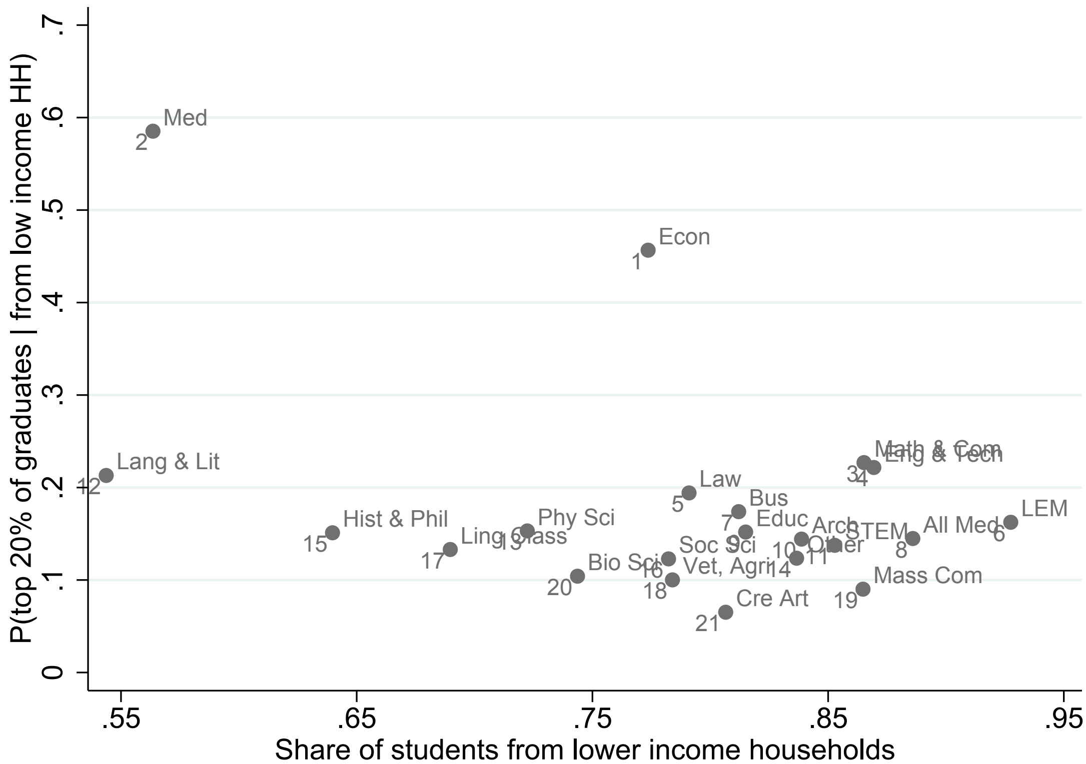


Figure 10: Subject mobility scorecard for women. Uses earnings from 2011/12-2013/14 tax years, treating individual observations as independent. The results are not very sensitive to this approach, however. The numbers represent within-gender subject ranking by the mobility index, which is the product of the x and y axes. See Appendix for full subject definitions. Numbers behind this figure are available in the Appendix.

Figures 11 and 12 repeat the same exercise, but for universities. This gives an indication of how each university is delivering individuals that come from lower income households into the top quintile of the graduate earnings distribution. The results show a clear negative relationship between the share of poorer students and the probability of them getting into the top 20% of the earnings distributions. The best performing of the named institutions are clearly those based in London, with the prominent universities of the LSE, Kings, Imperial and UCL all performing well for both genders. Of course this is partly because all students from these institutions are far more likely to be in the top quintile of the earnings distribution irrespective of a student's family background.

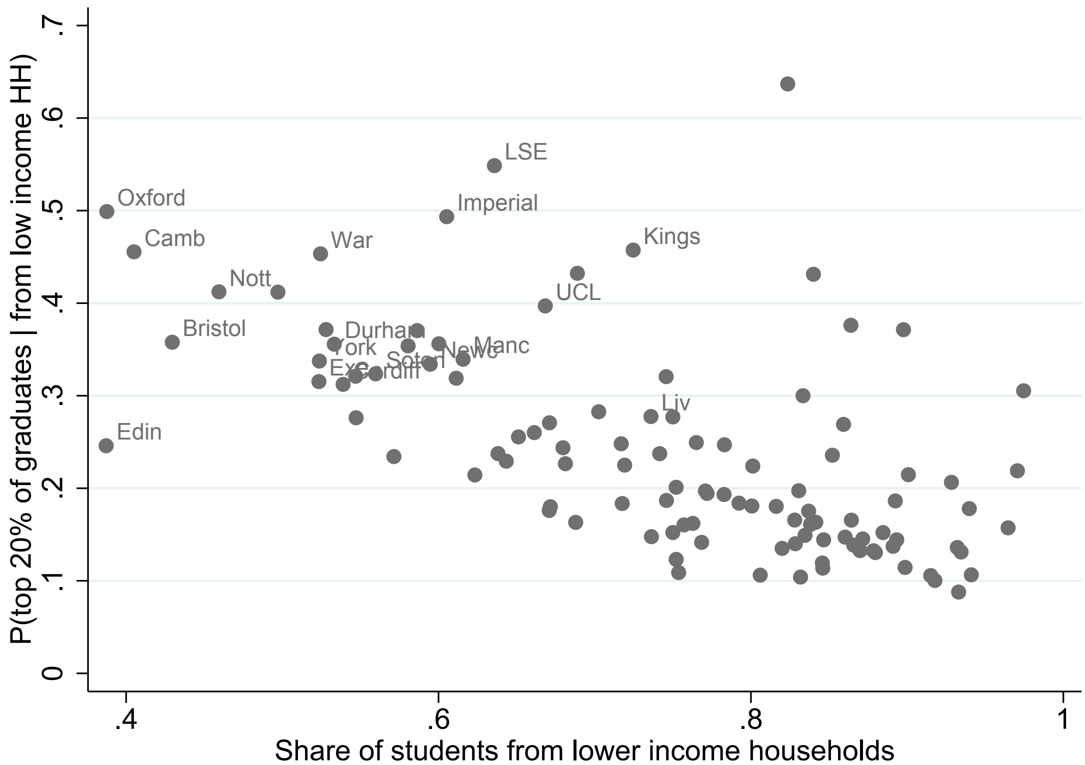


Figure 11: Institution mobility scorecard for men. Earnings rankings use the 2011/12-2013/14 tax years, treating individual observations as independent. The results are not very sensitive to this approach, however. We label a subset of universities we have been granted permission to name. Numbers behind this figure are available in the Appendix.

For men, around 60% of students at the prominent London-based institutions are from lower-income households, compared to less than 50% at Oxford, Cambridge and Bristol. The former all deliver at least 40% of these individuals into the top 20% of the earnings distribution, with the LSE doing the best out of the named institutions by delivering more than 50% into the top. Warwick is the highest performing of the named non-London institution, ranking 13th overall. It accepts similar shares of poor students to Durham, York, Exeter, Southampton and Cardiff, but is considerably more successful at delivering them to the top of the earning distribution. The worse performing institutions have a delivery rate of under 10%.

For women, the LSE and Imperial College are the stand-out performers, accepting similar shares of poorer students to Newcastle, Manchester, UCL, Southampton and Liverpool, but performing dramatically better in terms of delivery into the top. Oxford, Cambridge and Bristol are again similar, with amongst the lowest shares of students from poorer backgrounds and delivery in to the top of around 30%. Manchester is the best performing named non-London institution.

The above estimates, as with the earlier regression analyses, are descriptive. The results may not therefore entirely reflect the causal impact of these institutions on students' earnings. There may be selection effects and in some cases location effects from being near the high wage London labour market. However, they do illustrate the point that some institutions admit a large number of lower income students but such students do not necessarily go on to have high earnings, whereas some institutions admit far fewer but are more successful in delivering such students into the top of the income distribution.

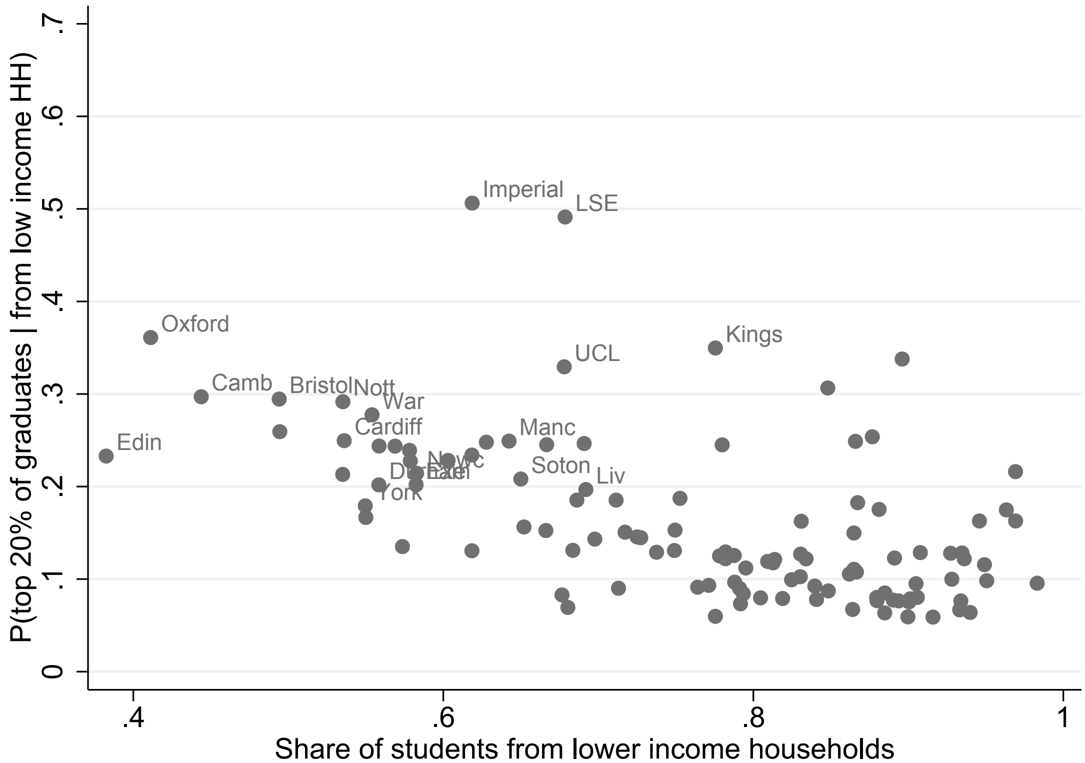


Figure 12: Institution mobility scorecard for women. Earnings rankings use the 2011/12-2013/14 tax years, treating individual observations as independent. The results are not very sensitive to this approach, however. We label a subset of universities we have been granted permission to name. Numbers behind this figure are available in the Appendix.

6 Conclusions

Using an innovative administrative data set, consisting of hard linked tax and student loan individual level data, as well as aggregate data on the graduates' degree course, we document how the earnings of graduates from higher and lower income households vary even after allowing for differences in subject taken and institution attended. The paper is the first of its kind to use such data in the English context to examine the correlation between a measure of parental income and graduates' earnings whilst being able to take account in some detail for the type of higher education experienced.

A main finding from this paper is that graduates' family background - specifically whether they come from a lower or higher income household - continues to influence graduates' earnings long after graduation. The socio economic gap in graduates' earnings is by no means entirely explained by differences in the subjects studied or institutions attended, though it is approximately halved once we account for these factors. When we take account of different student characteristics, degree subject and institution attended, the gap between graduates from higher and lower income households is still sizeable, at around 10% at the median. Further, we find that the gap is larger at the 20th and 90th percentiles of the graduate earnings distribution, suggesting that coming from a higher income household both protects against low earnings and provides greater opportunity for very high earnings. The magnitude of this effect is sufficient to be important and raises questions about the extent to which higher education can ensure that the labour market prospects of students from lower and higher income backgrounds are similar.

There are several possible explanations for our results. Students from wealthier families may have greater financial support from parents, may be more likely to relocate for work and may also be able to take greater career and financial risks than students from poorer backgrounds. They are likely to have better access to financial, social and cultural capital. Graduates' from lower income backgrounds may experience overt or covert discrimination in the labour market that constrains their earnings. All these factors will influence their career prospects and earnings and may explain why there continues to be a gap in the earnings of students from rich and poor backgrounds even after they experience the same higher education. An alternative explanation that we cannot completely discount is that there are some unobserved characteristics of students from higher income households that are correlated with their degree success and their earnings (e.g. social confidence).

Our results differ from previous work examining this issue in the US, namely Chetty et al. (2017) which finds that amongst US higher education graduates, family income does not appear to influence subsequent earnings, conditional on university choice. This is an important difference that is unlikely to be explained by methodological issues. Our measure of parental income does differ from theirs, as discussed, but we believe that our estimate is, if anything, an underestimate of the true earnings differences. Our dataset also enables us to allow for subject of degree and theirs does not. However, this too is unlikely to explain the difference in results unless on average US students from poor backgrounds tend to choose subjects that attract much higher earnings. Our research therefore provides preliminary evidence that the English labour market differs from that of the US, in that the advantages conferred by coming from a family with higher income (better social capital and networks, greater investment in non-cognitive skill, the ability to take risks etc.) are greater in the UK context. Since we cannot identify causal mechanisms, future work would usefully focus on understanding how it is that family income continues to impact on students' success in the English labour market even after higher education. There is a need for further research, not least to inform universities about any role they might play in assisting graduates from poor backgrounds to make the transition into the labour market and firms who may wish to change recruitment strategies that advantage students from higher income homes, such as unpaid internships.

References

Altonji, J. G., P. Arcidiacono, and A. Maurel (2015). The analysis of field choice in college and graduate school: Determinants and wage effects. Technical report, National Bureau of Economic Research.

Blanden, J., P. Gregg, and L. Macmillan (2007). Accounting for intergenerational income persistence: Noncognitive skills, ability and education. *Economic Journal 117*, C43–C60.

Blundell, R., L. Dearden, and B. Sianesi (2005). Evaluating the effect of education on earnings: Models, methods and results from the national child development survey. *Journal of the Royal Statistical Society, Series A 168*, 473–513.

Bratti, M., R. Naylor, and J. Smith (2005). Variations in the wage returns to a first degree: Evidence from the British Cohort Study 1970.

Britton, J., L. Dearden, N. Shephard, and A. Vignoles (2016). How English domiciled graduate earnings vary with gender, institution attended, subject and socio-economic background. Technical report, Institute for Fiscal Studies WP1606.

Britton, J., N. Shephard, and A. Vignoles (2015). Comparing sample survey measures of English earnings of graduates with administrative data. Unpublished paper: Department of Economics, Harvard University.

Bukodi, E. and J. Goldthorpe (2011a). Social class returns to higher education: Chances of access to the professional and managerial salaries for men in three British birth cohorts. *Longitudinal and Life Course Studies 2*, 1–71.

Bukodi, E. and J. H. Goldthorpe (2011b). Class origins, education and occupational attainment in Britain. *European Societies 13(3)*, 347–375.

Chetty, R., J. Friedman, E. Saez, N. Turner, and D. Yagan (2017). Mobility report cards: The role of colleges in intergenerational mobility. Technical report, mimeo, Stanford University.

Chevalier, A. (2011). Subject choice and earnings of UK graduates. *Economics of Education Review 30*, 1187–1201.

Chowdry, H., C. Crawford, L. Dearden, A. Goodman, and A. Vignoles (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.

Crawford, C., L. Dearden, J. Micklewright, and A. Vignoles (2016). *Family Background and University Success: Differences in Higher Education Access and Outcomes in England.* Oxford University Press.

Crawford, C. and A. Vignoles (2014). Heterogeneity in graduate earnings by socio-economic background. Unpublished paper: Institute of Fiscal Studies.

Dale, S. B. and A. B. Krueger (2014). Estimating the effects of college characteristics over the career using administrative earnings data. *Journal of Human Resources 49*(2), 323–358.

Dolton, P. and A. Vignoles (2000). The incidence and effects of overeducation in the UK graduate labour market. *Economics of Education Review 19*, 179–198.

Green, F., S. Machin, R. Murphy, and Y. Zhu (2012). The changing economic advantage from private schools. *Economica 79*, 658–679.

Hussain, I., S. McNally, and S. Telhaj (2009). University quality and graduate wages in the UK. CEE Discussion Paper No. 99.

Macmillan, L., C. Tyler, and A. Vignoles (2013). Who gets the top jobs? The role of family background and networks in recent graduates access to high-status professions. *Journal of Social Policy*, 1–29.

Naylor, R. (2002). Sheer class? The extent and sources of variation in the UK graduate earnings premium. LSE STICERD Research Paper No. CASE054.

Sloane, P. J. and N. C. O'Leary (2005). The return to a university education in Great Britain. *National Institute Economic Review*, 75–89.

Smith, J. and R. A. Naylor (2001). Determinants of individual degree performance: Evidence for the 1993 UK university graduate population from the USR. *Oxford Bulletin of Economics and Statistics 63*, 29–60.

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

Walker, I. and Y. Zhu (2013). The impact of university degrees on the lifecycle of earnings: Some further analysis. *BIS Research Paper No.112*.

Appendix A: Voluntary Repayments

	Repayments [1]	Add Gender [2]	Add Earnings [3]	Add HESA [4]
High Borrowing (Type B)	-837.6***	-817.5***	-826.4***	-505.3***
	(61.4)	(61.4)	(61.4)	(63.2)
Low Borrowing (Type A)	-1171.8***	-1155.6***	-1163.2***	-858.2***
	(66.3)	(66.3)	(66.3)	(75.0)
Female		-387.0***	-397.3***	-333.4***
		(51.3)	(51.4)	(53.9)
Earnings			-0.004**	-0.008***
			(0.001)	(0.001)
Constant	3598.9***	3807.0***	3913.3***	3167.9***
	(83.2)	(87.6)	(95.0)	(720.6)
N	22,176	22,176	22,176	22,176
Adjusted R^2	0.068	0.071	0.071	0.096

Table A1: Size of total voluntary repayments (£), conditional on making them, relative to Type X individuals. *** indicates significant at the 1% level; ** the 5% level. Female is a dummy set equal to one for women. Controls for cohort, age and year are included in all columns.

Appendix B: Heterogeneity estimates

	No Intake Information [1]	Bottom HEPs [2]	Lower Middle HEPs [3]	Middle HEPs [4]	Top HEPs [5]	Very Top HEPs [6]
High income HH	0.0982***	0.0993***	0.0945***	0.130***	0.0951***	0.256***
	(0.016)	(0.009)	(0.007)	(0.007)	(0.013)	(0.023)
Cohort poly	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Subject controls	Yes	Yes	Yes	Yes	Yes	Yes
HESA controls	Yes	Yes	Yes	Yes	Yes	Yes
HEI controls	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.095	0.080	0.105	0.155	0.177	0.182
N	79,500	89,436	99,279	94,106	25,875	11,267

Table A2: Estimation of % earnings differences between high and low income households by HEP group, for men. Includes the 1999-2005 cohorts and earnings data between 2008/09 and 2013/14. *** indicates significant at the 1 percent level. Standard errors are clustered at the individual level. Table provides raw data behind Figure 6 in main text.

	No Intake Information [1]	Bottom HEPs [2]	Lower Middle HEPs [3]	Middle HEPs [4]	Top HEPs [5]	Very Top HEPs [6]
High income HH	0.158***	0.104***	0.0913***	0.0986***	0.108***	0.0671**
	(0.014)	(0.008)	(0.006)	(0.006)	(0.012)	(0.023)
Cohort poly	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Subject controls	Yes	Yes	Yes	Yes	Yes	Yes
HESA controls	Yes	Yes	Yes	Yes	Yes	Yes
HEI controls	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.079	0.057	0.073	0.109	0.140	0.144
N	88,005	114,069	120,243	111,085	26,750	10,120

Table A3: Estimation of % earnings differences between high and low income households by HEP group, for women. Includes the 1999-2005 cohorts and earnings data between 2008/09 and 2013/14. *** indicates significant at the 1 percent level. Standard errors are clustered at the individual level. Table provides raw data behind Figure 6 in main text.

	1999 [1]	2000 [2]	2001 [3]	2002 [4]	2003 [5]	2004 [6]	2005 [7]
High income HH	0.141***	0.123***	0.137***	0.107***	0.117***	0.106***	0.0797***
	(0.012)	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)	(0.011)
Cohort poly	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subject controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
HESA controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
HEI controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.108	0.098	0.096	0.083	0.081	0.099	0.141
N	54,291	55,965	57,566	57,702	57,855	56,481	59,603

Table A4: Estimation of % earnings differences between high and low income households by cohort, for men. Includes earnings data between 2008/09 and 2013/14. *** indicates significant at the 1 percent level. Standard errors are clustered at the individual level. Table provides raw data behind Figure 7 in main text.

	1999 [1]	2000 [2]	2001 [3]	2002 [4]	2003 [5]	2004 [6]	2005 [7]
High income HH	0.118***	0.123***	0.144***	0.116***	0.0952***	0.0845***	0.0443***
	(0.011)	(0.009)	(0.009)	(0.009)	(0.009)	(0.010)	(0.009)
Cohort poly	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subject controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
HESA controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
HEI controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.086	0.080	0.069	0.063	0.070	0.071	0.130
N	61,806	65,101	67,071	67,607	68,203	68,198	72,286

Table A5: Estimation of % earnings differences between high and low income households by cohort, for women. Includes earnings data between 2008/09 and 2013/14. *** indicates significant at the 1 percent level. Standard errors are clustered at the individual level. Table provides raw data behind Figure 7 in main text.

	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14
	[1]	[2]	[3]	[4]	[5]	[6]
High income HH	0.0515***	0.0857***	0.113***	0.124***	0.133***	0.139***
	(0.010)	(0.011)	(0.010)	(0.009)	(0.009)	(0.009)
Cohort poly	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Subject controls	Yes	Yes	Yes	Yes	Yes	Yes
HESA controls	Yes	Yes	Yes	Yes	Yes	Yes
HEI controls	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.184	0.128	0.105	0.102	0.104	0.107
N	67,244	66,009	66,633	67,106	66,441	66,030

Table A6: Estimation of % earnings differences between high and low income households by year, for men. Includes the 1999-2005 cohorts. *** indicates significant at the 1 percent level. Standard errors are clustered at the individual level. Table provides raw data behind Figure 8 in main text.

	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14
	[1]	[2]	[3]	[4]	[5]	[6]
High income HH	0.0589***	0.0837***	0.102***	0.110***	0.119***	0.111***
	(0.009)	(0.009)	(0.009)	(0.008)	(0.009)	(0.008)
Cohort poly	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Subject controls	Yes	Yes	Yes	Yes	Yes	Yes
HESA controls	Yes	Yes	Yes	Yes	Yes	Yes
HEI controls	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.177	0.106	0.080	0.075	0.072	0.076
N	79,851	78,008	78,744	78,639	77,575	77,455

Table A7: Estimation of % earnings differences between high and low income households by year, for women. Includes the 1999-2005 cohorts. *** indicates significant at the 1 percent level. Standard errors are clustered at the individual level. Table provides raw data behind Figure 8 in main text.

Appendix D: Employment probabilities

	Men			Women		
		Add	Add		Add	Add
	Employment	Subject	HESA	Employment	Subject	HESA
	[1]	[2]	[3]	[4]		
High Income HH	-.003	.010	.014*	.046***	.058***	.036***
	(.006)	(.006)	(.006)	(.006)	(.006)	(.006)
$P(earn > 0 \mid$ Low income HH)	.867	.867	.867	.873	.873	.873
$P(earn > 0 \mid$ High income HH)	.866	.869	.870	.882	.884	.880
N	460,656	460,656	460,656	538,908	538,908	538,908

Table A8: Probit regression predicting employed. Predicted probability gives the probability of those from lower income households having non-zero earnings. *** indicates significant at the 1% level; ** the 5% level. Controls for cohort, age and year are included in all columns.

Appendix E: Mobility scorecard tables

Subject abbreviations used in Figures 9 and 10.

Medicine and Dentistry (Med); Subjects allied to Medicine (All Med); Biological Sciences (Bio Sci); Veterinary Sciences, Agriculture and related subjects (Vet, Agri); Physical Sciences (Phy Sci); Mathematical and Computer Sciences (Math & Com); Engineering and Technologies (Eng & Tech); Architecture, Building and Planning (Arch); Social studies (Soc Sci); Economics (Econ);

Law (Law); Business and Administrative studies (Bus); Mass Communications and Documentation (Mass Comm); Linguistics, Classics and related subjects (Ling Class); European Languages, Literature and related subjects (Lang & Lit); Historical and Philosophical studies (Hist & Phil); Creative Arts and Design (Cre Art); Education (Educ); STEM, LEM and Other represent miscellaneous STEM, LEM and Other courses - where individuals are in classes that are too small for us to be given their fine subject grouping. In these cases, we simply get their broader subject grouping.

Rank	Subject	Lower Income Share (A)	Delivery to top 20% (B)	Mobility Score (A*B)	Mobility Score Standard error	N
1	Medicine & Dentistry	0.604	0.662	0.400	0.007	4284
2	Economics	0.594	0.460	0.273	0.008	3510
3	LEM	0.883	0.266	0.235	0.008	2772
4	Education	0.825	0.263	0.217	0.004	12399
5	Maths & Computing	0.801	0.266	0.213	0.002	28071
6	Engineering & Technology	0.727	0.285	0.207	0.003	17916
7	Allied Medicine	0.846	0.241	0.204	0.005	7011
8	Law	0.771	0.259	0.200	0.004	7950
9	Business	0.776	0.234	0.182	0.002	26262
10	STEM	0.858	0.207	0.178	0.005	6708
11	Architecture	0.793	0.218	0.173	0.006	3993
12	Physical Sciences	0.666	0.228	0.152	0.003	11358
13	Languages & Literature	0.625	0.242	0.151	0.009	1776
14	Social Sciences	0.701	0.211	0.148	0.003	12762
15	Vetinary/Agriculture	0.833	0.178	0.148	0.009	1635
16	Other	0.827	0.173	0.143	0.002	24177
17	Linguistics & Classics	0.729	0.171	0.125	0.005	5094
18	History & Philosophy	0.601	0.202	0.122	0.004	8451
19	Biological Sciences	0.749	0.150	0.112	0.003	13947
20	Mass Communication	0.846	0.115	0.097	0.004	6897
21	Creative Arts	0.826	0.096	0.080	0.002	23355

Table A9: Subject mobility scorecard table for men.

Rank	Subject	Lower Income Share (A)	Delivery to top 20% (B)	Mobility Score (A*B)	Mobility Score Standard error	N
1	Economics	0.774	0.457	0.353	0.014	1206
2	Medicine & Dentistry	0.564	0.585	0.330	0.006	5994
3	Maths & Computing	0.865	0.227	0.196	0.005	7389
4	Engineering & Technology	0.869	0.222	0.193	0.007	2802
5	Law	0.791	0.194	0.154	0.003	14052
6	LEM	0.927	0.162	0.151	0.008	2151
7	Business	0.812	0.174	0.141	0.002	23880
8	Allied Medicine	0.886	0.145	0.128	0.002	19698
9	Education	0.815	0.152	0.124	0.002	35886
10	Architecture	0.839	0.144	0.121	0.010	1116
11	STEM	0.853	0.138	0.117	0.005	4749
12	Languages & Literature	0.544	0.213	0.116	0.005	3945
13	Physical Sciences	0.722	0.153	0.111	0.004	7218
14	Other	0.837	0.124	0.103	0.002	33678
15	History & Philosophy	0.640	0.151	0.097	0.003	8319
16	Social Sciences	0.782	0.123	0.096	0.002	20229
17	Linguistics & Classics	0.690	0.133	0.092	0.003	11862
18	Vetinary, Aggriculture	0.784	0.100	0.078	0.005	3402
19	Mass Communication	0.865	0.090	0.078	0.003	9984
20	Biological Sciences	0.744	0.104	0.077	0.002	21453
21	Creative Arts	0.807	0.065	0.053	0.001	30441

Table A10: Subject mobility scorecard table for women.

Institution	Lower Income Share (A)	Delivery to top 20% (B)	Mobility Score (A*B)	Mobility Score Standard error	N
Bristol	0.430	0.358	0.154	0.008	2193
Cambridge	0.405	0.455	0.184	0.008	2385
Cardiff	0.539	0.312	0.168	0.009	1581
Durham	0.533	0.356	0.190	0.008	2436
Edinburgh	0.387	0.246	0.095	0.010	798
Exeter	0.523	0.315	0.165	0.008	2115
Imperial	0.605	0.493	0.299	0.012	1383
King's College	0.725	0.457	0.331	0.012	1536
LSE	0.636	0.549	0.349	0.022	453
Liverpool	0.736	0.277	0.204	0.008	2409
Manchester	0.616	0.339	0.209	0.006	4278
Newcastle	0.595	0.334	0.199	0.008	2487
Nottingham	0.459	0.412	0.189	0.007	3585
Oxford	0.388	0.499	0.193	0.008	2523
Southampton	0.560	0.324	0.181	0.008	2562
UCL	0.668	0.397	0.265	0.010	1809
Warwick	0.524	0.453	0.238	0.009	2145
York	0.524	0.337	0.177	0.010	1392

Table A11: Institution mobility scorecard table for men. Shows named institutions only, in alphabetical order.

Institution	Lower Income Share (A)	Delivery to top 20% (B)	Mobility Score (A*B)	Mobility Score Standard error	N
Bristol	0.494	0.295	0.146	0.007	2295
Cambridge	0.444	0.297	0.132	0.007	2670
Cardiff	0.536	0.250	0.134	0.007	2205
Durham	0.558	0.202	0.113	0.006	2697
Edinburgh	0.382	0.233	0.089	0.008	1302
Exeter	0.582	0.202	0.118	0.007	2220
Imperial	0.619	0.506	0.313	0.018	645
King's College	0.776	0.350	0.271	0.009	2205
LSE	0.679	0.491	0.333	0.030	252
Liverpool	0.692	0.197	0.136	0.007	2688
Manchester	0.642	0.249	0.160	0.005	4530
Newcastle	0.583	0.214	0.125	0.007	2538
Nottingham	0.535	0.292	0.156	0.006	3408
Oxford	0.411	0.361	0.148	0.008	2196
Southampton	0.650	0.208	0.135	0.006	3060
UCL	0.678	0.329	0.223	0.010	1854
Warwick	0.554	0.278	0.154	0.007	2361
York	0.550	0.179	0.098	0.008	1239

Table A12: Institution mobility scorecard table for women. Shows named institutions only, in alphabetical order.