I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12826

Is 'First in Family' a Good Indicator for Widening University Participation?

Anna Adamecz-Völgyi
Morag Henderson
Nikki Shure

DECEMBER 2019

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12826

Is 'First in Family' a Good Indicator for Widening University Participation?

Anna Adamecz-Völgyi
University College London and KRTK

Morag Henderson
University College London

Nikki Shure
University College London and IZA

DECEMBER 2019

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Is 'First in Family' a Good Indicator for Widening University Participation?*

Universities use 'first in family' or 'first generation' as an indicator to increase the diversity of their student intake, but little is known about whether it is a good indicator of disadvantage. We use nationally representative, longitudinal survey data linked to administrative data from England to provide the first comprehensive analysis of this measure. We employ parametric probability (logit) and non-parametric classification (random forest) models to look at its relative predictive power of university participation and graduation. We find that being first in family is an important barrier to university participation and graduation, over and above other sources of disadvantage. This association seems to operate through the channel of early educational attainment. Our findings indicate that the first in family indicator could be key in efforts to widen participation at universities.

JEL Classification: I23, I24, J24

Keywords: socioeconomic gaps, higher education, widening participation, first in family, first generation, educational mobility, machine learning, predictive models

Corresponding author:
Nikki Shure
UCL Institute of Education
Department of Social Science
20 Bedford Way
London, WC1H 0AL
United Kingdom

E-mail: nikki.shure@ucl.ac.uk

* The Nuffield Foundation is an endowed charitable trust that aims to improve social wellbeing in the widest sense. It funds research and innovation in education and social policy and also works to build capacity in education, science and social science research. The Nuffield Foundation has funded this project (grant number EDO/43570), but the views expressed are those of the authors and not necessarily those of the Foundation. More information is available at www.nuffieldfoundation.org. We thank John Jerrim, Charles Livesey, George Ploubidis, Richard Silverwood, Anna Vignoles, the participants of the IE HAS Seminar, the Budapest Institute Seminar, ESPE annual conference, LSE CEP Education Work In-progress Seminar, and UCL Widening Participation Research Seminar for helpful comments.

1. Introduction

A body of literature has established the importance of higher education (HE) in achieving social mobility (Blanden and Macmillan 2016; Chetty et al. 2014) and a range of positive life outcomes (Oreopoulos and Petronijevic 2013). At the same time, there is a large evidence base that university participation and graduation is graded by socioeconomic status, with individuals from advantaged backgrounds more likely to attend and complete university than their peers from disadvantaged backgrounds (Blanden and Machin 2004; Britton et al. 2016; Walker and Zhu 2018). In the interest of improving fairness and social justice, universities across the world have increased their efforts to diversify the profile of their student intake. This means focusing on disadvantaged individuals who were traditionally less likely to attend or complete university. The first step in this process is identifying individuals from disadvantaged backgrounds using a range of disadvantages and the indicators that can be used to measure them. As with any indicators, practitioners must weigh the administrative ease of collection with validity and accuracy.

In England, the effort to increase the diversity of student intake is known as the 'Widening Participation' (WP) agenda. The WP agenda focuses on increasing access and participation from disadvantaged and vulnerable groups, including those from low income families, those who are eligible for Free School Meals (FSM), those from a low social class background, those who could be 'first in family' or 'first generation' graduates, those who were young carers or who have been in care, those with disabilities or special education needs (SEN), and those from minority ethnic backgrounds. In order to attract students from these groups, universities need to use a variety of indicators to capture the aforementioned disadvantages and be able to weigh the strengths and weaknesses of the indicators they use. Currently, there is limited evidence available upon which universities can base their decision making.

According to the existing empirical literature, the most important factor of HE success is prior educational attainment (Gorard et al. 2017). As socially and economically disadvantaged students lag behind in their educational attainment well before the time of university entry (Chowdry et al. 2013), prior educational attainment already captures some of the negative effects of social and economic hardship on HE participation and graduation. It remains challenging to identify whether the same background characteristics that negatively affect pre-university achievement have any remaining significant impact on HE participation and graduation after controlling for the effect of prior educational attainment, because prior educational attainment is the consequence of the same background characteristics. The usual approach to this question is to estimate the correlational relationship between HE participation (or graduation) and the various measures of disadvantage in a regression framework in a sequential fashion: first including indicators of socioeconomic disadvantage as explanatory variables, such as parental social background or income, and then extending them with prior educational attainment to look at whether the initially estimated coefficients of background variables change. Using this method, Chowdry et al. (2013) find that although taking prior educational attainment into account reduces the size of the negative effects, i.e. most of the negative effect of low socioeconomic background operates through pre-university attainment, low socioeconomic status (SES) still significantly reduces the probability of HE participation even after controlling for prior educational attainment. Similarly, Vignoles and Powdthavee (2009) and Crawford (2014) find that low social background increases the probability of dropping-out among university students even after controlling for prior educational attainment.

In this paper, we provide the first robust analysis of a range of WP indicators and ask whether 'first in family' is a good indicator for widening participation. In this context, 'first in

family' (FiF) graduates refers to students who attend university and achieve a university degree but whose (step) mother and (step) father did not. We use the term 'potential FiF' to describe the indicator that identifies young people who could be the first in their family to achieve a university degree because neither of their parents has one. If accepted by a university, they will become FiF university attendees and if they graduate, they will become FiF graduates. We focus on FiF for four main reasons: (1) a large proportion of the English population come from families where no parent has a university degree; (2) while most other WP measures target the very bottom of the socioeconomic distribution (usually the lowest quintile in terms of parental income-related measures), potential FiF might be able to capture the relatively less advantaged students across the distribution; (3) a large range of universities and employers currently use it as a measure; and (4) it is the only WP measure that directly captures an individual's parental human capital.

While there is substantial research on other forms of socioeconomic disadvantage limiting access to higher education (Boliver 2013; Bowen, Chingos, and McPherson 2009; Chowdry et al. 2013), Henderson, Shure, and Adamecz-Volgyi (2019) provide the first descriptive results on FiF young people in England. Unlike some other sources, i.e. the Higher Education Statistics Agency (HESA) in the UK, they define (step) parents holding a university degree (BA/BSc or higher) as graduates, and do not consider below-degree level HE certificates and diplomas as graduation. Using this definition, they show that 84 percent of a recent cohort of young people born in England in 1989 could potentially be first in family since none of their parents hold a university degree. Those who go on to attend and graduate as the first in their family from university make up 18 percent of the overall cohort and comprise two-thirds of the cohort's university graduates (Henderson, Shure, and Adamecz-Volgyi 2019). This seemingly large proportion of FiF graduates is comparable to similar statistics from administrative data (see the Data section and Appendix A for more detail) and importantly applies to a large swath of the population, unlike many of the other WP measures, a point we return to in the discussion of the paper. Henderson, Shure, and Adamecz-Volgyi (2019) show that FiF graduates tend to come from ethnic minority and lower socioeconomic backgrounds as compared to their peers who match their parents with a degree but come from slightly more advantaged backgrounds and already outperform their peers who match their parents without a degree during compulsory education.

To the best of our knowledge, the majority of Russell Group universities (see the Appendix in Henderson, Shure, and Adamecz-Volgyi (2019)) currently use potential FiF as an explicit characteristic in order to widen participation in addition to a range of other universities and employers.[1] It is likely that an even higher proportion will use potential FiF when introducing contextual admissions[2] (CA) procedures (Boliver et al. 2017), increasing the need for robust evidence on this measure. Currently, potential FiF is treated by universities as a measure of disadvantage, which may manifest in terms of impeded aspirations and expectations, lack of understanding of higher education and the labour market, and lower parental human capital. It may, however, also be considered as a signal for upward educational mobility and ambition. Therefore, understanding this 'disadvantaged' group relative to other social disadvantages including low household income, disability, and status disadvantage is important.

[1] At the time when this cohort was applying to university in 2007/08, a range of universities would have used FiF alongside other WP indicators. As the former head of the Higher Education Funding Council for England (HEFCE), John Selby, points out, universities treatment of "disadvantaged backgrounds [was] (variously measured)", which makes it challenging to provide an exact number (Selby 2018).

[2] Widening Participation policy refers to a broad set of measures to increase the share of students from differing socioeconomic backgrounds to attend HE, while Contextual Admissions refer to one particular tool of WP when universities and colleges make contextual [or lower] offers to students with particular sociodemographic characteristics in order to widen access.

Theoretically, potential FiF is a reasonable criterion to use as a WP measure because a broad literature documents the generally positive, although usually small, causal effects of parental education on children's schooling outcomes (see a review in Holmlund, Lindahl, and Plug (2011)). Although we know less about the causal effects of parental HE graduation in particular, the literature agrees that children of university graduates enjoy advantages in several domains. Currie and Moretti (2003) find that maternal higher education increases infant health, Suhonen and Karhunen (2017) reveal large and positive effects of parental graduation on children's university participation and years of schooling, while Maurin and McNally (2008) show that children of university graduates are less likely to repeat grades during their studies compared to children of parents who have never been to university. Thus, empirical evidence suggests that children of parents with university degrees are pre-determined to have higher education (and life) outcomes, and potential FiF students might lack this advantage.

Another strand of the literature investigates the best WP measures to use in the UK context.[3] None of these papers, however, have explicitly looked at potential FiF. Boliver, Gorard, and Siddiqui (2015) review WP measures used by universities as Contextual Admissions (CA) indicators. They state the most important theoretical criteria for WP measures are: reliability; measurability; validity; and accessibility. They argue the measures should have a clear (legal) definition; should not be missing systematically for any subgroups of students; should not create unintended inequalities; should be comparable across time and cohorts and should reflect individual circumstances rather than school- or neighbourhood-level information.

In follow-up research, Gorard, Siddiqui, and Boliver (2017) use a data-driven approach to examine the quality and the overlap of various WP measures. They look at the frequency and structure of missing values of WP indicators in the National Pupil Database and investigate their overlap and relationship with educational attainment using cross-tabulation. Due to quality problems and/or low correlation with educational attainment, they find IDACI scores, school type (private/public) and educational attainment to be less valuable and instead suggest using the following characteristics: gender (being male); month of birth; and FSM eligibility, as a measure of poverty. They also find having special educational needs (SEN) promising and living in care as safe to use.

Ilie, Sutherland, and Vignoles (2017) explore the qualities of FSM eligibility in terms of its overlap with other measures of disadvantage and their relative predictive power of educational attainment. They apply a multilevel linear modelling approach to predict Key Stage 4 (GSCE) examination results and compare other measures of low socioeconomic background with FSM eligibility in terms of explaining the within-school variation of educational attainment of students. They find that detailed measures of parental occupation (measured in eight categories), parental education (measured in seven categories) and some household characteristics (housing tenure, age of mother, household size) have slightly better explanatory power than FSM eligibility. Neighbourhood-level deprivation measures (IDACI score, average occupation) and household income perform slightly worse. When they include all these measures in a model together, the estimated coefficient of FSM eligibility is smaller but still significantly negative (18 instead of 51 points decrease in total capped GCSE scores), which they interpret as FSM eligibility capturing some additional sources of disadvantage over and above other measures. Earlier work on the evaluation of FSM eligibility as an indicator draws similar conclusions in terms of its use as a proxy for family income (Hobbs and Vignoles 2010), and in comparison to various measures of disadvantage (Gorard 2012).

[3] A much broader and related question might be whether using any WP indicators (or CA policies) could make access to HE fairer in general; however, discussing this question is beyond the scope of this paper.

This paper examines potential first in family (i.e. having parents without university degrees) status in a similar manner as an indicator of disadvantage. We use longitudinal survey data linked to the census of administrative education data to ask whether or not potential FiF picks up additional disadvantage beyond other traditional WP indicators. We show that potential FiF covers a large proportion of the sample, making it appear like a measure of 'non-advantage', but that the majority of potential FiF graduates also face at least one additional disadvantage. We examine the correlations between existing WP measures and see how much potential FiF overlaps with other existing WP measures.

Our contribution to the literature in terms of the applied methodology is threefold. First, while the existing literature investigates WP measures estimating regression models and using either the size and significance of the estimated coefficients or the magnitude of R-squared estimates to draw conclusions, we employ a direct measure of predictive power, the area under the ROC curve (AUC), to investigate the relative predictive power of FiF. We start by looking at the AUC of WP indicators in a logit framework, comparing them to each other one-by-one. Then, we add the potential FiF measure on top of a baseline set of WP indicators to see if it increases the predictive power of logit models. The second applied methodological contribution of this paper comes from the fact that unlike any previous attempts, we use a logit-based Least Absolute Shrinkage and Selection Operator (Lasso) procedure to pin down an empirically optimal combination of WP indicators as the baseline set of measures to be extended by FiF. Third, we are the first in this literature to apply a non-parametric prediction strategy, a random forest algorithm, to look at the relative importance of WP measures and potential FiF in predicting HE participation and graduation. We also use this framework to investigate whether early educational attainment is the most important channel of the effects of potential FiF on HE success.

Our results show that the potential FiF measure does capture some additional disadvantage beyond other traditional WP indicators. We find striking results indicating that having parents with university degrees is a fundamental driver of an individual's HE participation and graduation. When we compare potential FiF to other WP indicators, it emerges as the most important measure until we condition on prior attainment and all measures end up equally predictive. We provide evidence that the effects of family background manifest in earlier educational attainment and conclude that our results support Contextual Admissions as a main element of WP. This research is the first to explore the validity of the potential FiF indicator in England using large scale, nationally representative data and will hopefully inform the Widening Participation agenda of universities across the UK. The rest of this paper is structured as follows. In section 2, we provide an overview of the data. In section 3, we present the empirical strategy and in section 4, our results. Section 5 concludes with a discussion.

2. Data and descriptive statistics

We use Next Steps (formerly the Longitudinal Study of Young People in England, LSYPE) which follows a cohort of young people born in 1989/1990. Next Steps began in 2004 when the sample

members were aged between 13 and 14 and comprises eight waves of data until age 25.[4] This cohort of young people can be linked with the National Pupil Database, allowing us to access their national school exam results.

Respondents of Next Steps were selected to be representative of young people in England using a stratified random sample of state and independent schools, with disproportionate sampling for deprived schools, i.e. those in the top quintile of schools in terms of the share of pupils eligible to Free School Meals (Department for Education 2011).[5] In deprived schools, students of minority ethnic backgrounds were over-sampled to provide a sufficient number of observations for analysis (Centre for Longitudinal Studies 2018). Design weights were constructed to take care of the oversampling of deprived schools and ethnic minority students within deprived schools using inverse probability weighting such that "*the school selection probabilities and the pupil selection probabilities ensured that within a deprivation stratum, all pupils within an ethnic group had an equal chance of selection*" (Department for Education 2011). Starting from wave 1, attrition weights are estimated by stratum to take care of the initial school-level non-compliance as well as individual attrition from the study. The weighting procedure differs by school type (independent vs. state schools) and takes into account both school-level and individual-level information. The final models to predict the probability of individual non-response differ in each wave, and the estimated probabilities are carried across waves as the study progresses.

Schools are the primary sampling units of Next Steps, then pupils within schools. The two-stage sampling design presents a possible clustering effect due to between-school differences and potential school-specific random shocks. Therefore, we control for school-specific variables and allow for the intra-school correlation of random shocks by clustering the standard errors by schools in our parametric models. In the first four waves, both young people and their parents were interviewed, and the information content of all variables on family background and parental education that we use in this paper was reported directly by the parents. From age 17, only young people were interviewed.

We make use of the Next Steps age 25 data, which follows up sample members as they develop into young adults with a particular focus on highest level of educational qualification achieved and characteristics of the university at which they studied. The age 25 wave of Next Steps covers 7,707 young people, 36.7% of the initially drawn sample and 49% of the actual sample of the first wave. All results that we present in this paper are estimated using the final weights to take care of initial oversampling of disadvantaged schools and ethnic minority students, school non-compliance, the ethnic boost added to the sample at age 16, and attrition in all waves. In order to avoid dropping cases with missing or unknown information on WP measures or background variables, we take the first available response mentioned for parental class, parental education and

[4] The timing of this cohort means that the young people were affected by increased tuition fees in higher education. In 2006-07, top-up university fees were introduced in England and Northern Ireland, and despite universities being allowed to choose their fee amount, almost all UK institutions chose to charge the full £3,000 per annum fee (Wyness 2010). In addition to this policy change, the Next Steps cohort also faced some administrative changes in loan and grant entitlement, which ultimately did not result in an overall change to access to finances, rather changes in the application process (see Wyness (2010) for additional information).

[5] Initially, 54 independent and 646 state-maintained schools were chosen, but almost half of the independent schools (especially those in inner-London) and a fifth of state schools decided not to participate. The first wave thus started with a 21,000-observation issued sample of 13/14-year old students in 28 independent and 646 maintained schools with an average response rate of 74%, resulting in a 15,770-observation initial sample. In wave 4, a 600-participant ethnic boost sample were added to the study, selected from the schools that were chosen at the beginning but did not cooperate in wave 1 (Centre for Longitudinal Studies 2018).

household tenure over the first four waves. We include missing flags to deal with remaining item non-response. Note that we tested that our main results hold on the complete case sample as well.

There is a discussion in the literature as to whether or not Next Steps is the most desirable data source to exploit when looking at HE outcomes in particular. Anders (2012) examines the quality of the information on HE participation by the seventh wave of Next Steps. He finds that Next Steps estimates HE participation by age 19/20 as 43.3%, which is about 10 percentage points larger than the Higher Education Initial Participation Rate (HEIPR)[6] calculated by the Department for Education for English domiciled young people aged 17-19 who spend at least six months in HE (Department for Education 2017). More recent Next Steps data indicates this may be less of a problem. In the eighth wave, collected in 2015, the share of HE participants and the HEIPR rate for those aged 17/25 are very similar (40.0% (weighted) in Next Steps, 40.4% HEIPR, Table 1).

Sample

In this paper, we think about FiF (and all WP measures) from the point of view of universities: when looking at a pool of applicants, should universities use FiF to widen participation? As we do not observe university applications in the data, we impose a sample restriction in order to pin down the potential pool of applicants for university, i.e. those with enough formal education. We restrict our analytical sample to those who theoretically would have had the possibility to apply and attend university, i.e. who decided to stay in school after completing compulsory exams at age 16. Formally, this means they started "Level 3 studies". The most common form of Level 3 studies is A levels, but individuals can also pursue vocational Level 3 studies that still enable them to attend university (Department for Education n.d.). An alternative sample restriction would be to include only those who completed their exams at the end of compulsory schooling with fairly good grades, usually captured in the literature as those who achieved at least 5 A-C grade GCSE examinations. Although this is not an official threshold for university application, students below this attainment level are unlikely to successfully apply to a university. The 5 A-C GCSE threshold is used for example by studies estimating returns to graduation (Belfield et al., 2018).

The main point of any sample restriction is to create somewhat comparable groups. Ideally, we would want to compare the HE outcomes of students who do not differ from each other in any other dimension other than potential FiF. Whatever sample restriction we use based on prior educational achievement, we leave out the lower part of the distribution of students in terms of their potential HE attainment. As children with lower-educated parents are likely to have lower educational attainment than children of higher-educated parents, those excluded are more likely to be FiF and less likely to go to university. Thus, this procedure, while helpful in creating more comparable groups, would also cause us to underestimate the negative statistical relationship between being FiF and HE outcomes. The 5 A-C GCSE-restriction selects *better* students based on their exam grades/abilities while the Level 3-restriction focuses more on motivation for schooling and shows students' revealed preferences for learning. As we explicitly control for prior

[6] The HEIPR is applicable to students who live in England, enter HE for the first time, attend UK HE institutions and English, Welsh and Scottish further education colleges, and, stay in HE for at least six months (Department for Education 2017). It is an estimate of the likelihood of a young person participating in HE, in our case, by age 25, i.e. the sum of the likelihoods of HE entry at ages 17-25. This is estimated using the administrative data of the Higher Education Statistics Agency on the number university participants and the size of each cohort from the UK Census (Office for National Statistics). As the yearly publications of the Department for Education cover the data of those at age 17-25 in each year, we constructed the HEIPR rate of our particular cohort (those born in 1989-90) by using the appropriate year of observation for all ages (i.e., the data of the 2014/15 academic year for age 25, the data of the 2013/14 academic year for age 24, …, etc.).

educational attainment as a proxy for cognitive abilities in our models, we chose to restrict the sample to those who attended Level 3 studies. Note, however, that we have conducted additional analysis and our main results hold on the total sample or on the 5 A-C GCSE sample as well. Out of the 7,707 individuals at age 25, 5,047 young people attended Level 3 studies. We use this subset of Next Steps for the main empirical analysis in this paper.

Outcome variables

We use two binary outcome variables: HE participation and HE graduation. Both variables are supplied as derived variables at age 25 based on information gathered from all waves and they have no missing values. University participation is defined broadly as "ever been to university", independent from the length of university attendance. University graduation is defined as whether a young person achieved a first degree/foundation degree (BA/BSc) by age 25. It does not cover any other types of higher education qualifications that are usually awarded after shorter-than-three-year higher education courses (diplomas, certificates, awards, etc.). As we have mentioned above, the share of HE participants is 40.0% in the Next Steps data, very similar to the HEIPR (Table 1).

<<<Table 1>>>

The weighted share of graduates in Next Steps is 26.9% while in the most comparable data source, the Annual Population Survey (APS), it is 39.6%. There are however significant differences between the two samples and the two definitions. The APS sample is representative of 80,000 households across the UK and therefore may not be representative of 25-year-olds, whilst Next Steps is representative of 25-year-olds who have been living in England since 2004. The APS graduation rate also takes all types of Level 4 degrees into account, while in Next Steps we only look at BA/BSc degrees or higher (and thus exclude Level 4 specifications below university degree level). Appendix A includes an additional comparison of Next Steps to HESA data, which also confirms the reliability of our outcome variables.

These comparisons with administrative data, highlighting key sampling differences, give us confidence that the measurement of our outcome variables, higher education participation and graduation, is accurate and reliable in this sample.

Widening Participation measures

We examine a range of commonly used WP measures to which we compare the first in family measure. A young person is a potential first in family if neither his/her biological mother or father (or resident step parents) had achieved a university degree (BA/BSc or higher) by the time the young person was aged 17, i.e. before university application. Although their parents may return to higher education once the young person is older than 17, we choose this point in order to examine the importance of growing up with parents without university degrees. Our focus is on intergenerational educational mobility as we are unable to look at whether a sibling attended university due to data constraints. This also corresponds with the Widening Participation indicator, which only asks whether or not an individual's (step) parents achieved a degree.

In order to identify the prevalence and overlaps of the indicators of socioeconomic disadvantage that are used by the Widening Participation agenda, we focus on 10 forms of disadvantage indicators currently applied by universities. As previously mentioned, there is substantial heterogeneity in how universities measure disadvantage (Selby 2018), which means we take a pragmatic approach: the focus on these 10 measures is the result of data availability, i.e. what

measures are available in Next Steps, and WP policy. We make use of the first five waves of Next Steps[7] to capture these measures. In particular, based on Boliver, Gorard, and Siddiqui (2015) and Ilie, Sutherland, and Vignoles (2017), we look at the following binary indicators:

1. SEN: whether the young person ever reported any special educational needs from age 13/14 to 16/17;
2. FSM: whether the young person ever reported to be eligible for Free School Meals from age 13/14 to 17/18;
3. Low social class: whether the highest parental social class[8] of the family mentioned from age 13/14 to 17/18 was "Routine occupations or not currently working";
4. Income deprivation: whether the family belonged to the top 20% of deprived families based on the Income Deprivation Affecting Children Index (IDACI)[9] measured in ages 14/15 and 15/16;
5. Young carer: whether the young person ever reported to regularly providing unpaid care to anyone in the household from age 13/14 to 15/16;
6. Non-White: belongs to a non-White ethnic group;
7. Living with disability (classified according to the Disability Classification Equality Act (2010)) or long-term illness that affects schooling at least once in ages 13/14 to 16/17;
8. Single HH: lived in a single-parent household at least once anytime at or before age 17/18;
9. Care leaver: whether the young person has ever been in care by age 13/14 or lived at least once with no parent in the household in ages 13/14 to 17/18;
10. Multiple deprivation: top 20% of the 2004 Index of Multiple Deprivation (IMD)[10] captured in ages 14/15 to 15/16.

It is important to note that IDACI and IMD are measured at the neighbourhood level and linked to the individual using postcode while all other measures of disadvantage are measured at the household or individual level.

Control variables

In terms of the individual characteristics of young people, we control for gender and educational attainment captured by Key Stage 2 (KS2) total score measured at age 11 and capped linear GCSE

[7] Next Steps waves cover the sample in the following ages: wave 1, age 13/14: wave 2, age 14/15; wage 3, age 15/16; wave 4, age 16/17; wave 5, age 17/18; wave 6, age 18/19; wave 7, age 19/20; and wave 8, age 25/26.

[8] Social class is measured using the National Statistics Socio Economic Classification (NS-SEC) which uses occupational types to capture dimensions of social class (Office for National Statistics n.d.; Rose and Pevalin 2001).

[9] The Income Deprivation Affecting Children Index (IDACI) score is an index which represents the proportion of children under the age of 16 living in a low-income household by Lower-Layer Super Output Area (LSOA) (Department for Communities and Local Government 2015). These scores were computed into quintiles and a binary measure was created which shows the most deprived 20 percent of neighbourhoods in England.

[10] The Index of Multiple Deprivation (IMD) is the official measure of relative deprivation for small areas in England constructed by the Department for Communities and Local Government. The IMD ranks Lower-Layer Super Output Areas (LSOA) based on seven domains of deprivation (income, employment, education, health, crime, barriers to housing, living environment) from least to most deprived (Department for Communities and Local Government 2015).

(Key Stage 4) score measured at age 16.[11] In some models, we also use Key Stage 3 (KS3) total scores measured at age 14. We create categorical variables based on the quintiles of all of these measures, which allows us to construct a sixth category for missing values due to linkage issues between Next Steps and the National Pupil Database.

Descriptive statistics

We begin our analysis of potential FiF and the other WP indicators by assessing their prevalence in the sample. This is shown in Table 2. Having parents with no degree, i.e. being potentially FiF, is the most common disadvantage: over 75 percent of those with Level 3 education are potential FiF. This means that being a potential FiF should perhaps be viewed as a measure of "non-advantage" as opposed to disadvantage since it is the by far the most prevalent WP indicator and applies to the majority of this cohort. It is also not the case that being a potential FiF is a non-advantage that occurs in isolation. Table 2 shows that 81.6% of the potential FiF face at least one additional disadvantage. This proportion is lower, however, than other WP indicators, where the share of those facing at least one more disadvantage is between 91.7-100% (Table 2, third column). In other words, almost one-fifth of the potential FiF face no any other types of disadvantage; they could be viewed as not "really" disadvantaged in practice, but of course may still face barriers in terms of lack of role models and information about HE. The potential FiF group may be disadvantaged in a way not captured by the other WP measures. Taken together, this indicates that the potential FiF group is heterogeneous in nature.

<<<Table 2>>>

Next, we examine the overlap of these WP indicators by calculating the proportion of our sample facing multiple disadvantages. Figure 1 shows that only 7.9 percent of those with Level 3 education face no form of social disadvantages; 22.7% of the sample face one, 26.1% face two, and more than 40% face three or more types of disadvantage at the same time. This shows that while disadvantages cluster in some individuals, a large proportion of the population is disadvantaged according to at least one of these measures.

<<<Figure 1>>>

Table 3 captures the proportion of potential FiF across all WP measures. In particular, it compares the share of potential FiF among those facing (column 1) and not facing the given type of disadvantage (column 2). The difference of shares (column 3) is then tested via two-sided t-tests for statistical significance. The share of potential FiF is about the same among those who have had or not have had special education needs, who are non-White, disabled or care leavers. The share of potential FiF is however much larger among those eligible for FSM (+15.8 percentage points (pps)), coming from low social class backgrounds (+15.3 pps), suffering from multiple deprivation (+14.9 pps) or income deprivation (+14.5 pps), young carers (+7.2 pps) and single parent households (+9.8 pps). It seems that simply looking at the raw data, potential FiF status is

[11] GCSEs are two-year courses which are taken when an individual is aged 14-16 and are considered Key Stage 4 qualifications. This linear measure is created by assigning values to the grades, for example, Grade G is worth 16 points. Each grade improvement thereafter, for example, from G to F, C to B, or A to A*, is equivalent to an additional six points. The capped linear score, takes the best eight GCSE subjects scores. This measure is standardised as it takes into account the fact that students may take a different number of GCSEs (and resits) and enables better comparability than the total GCSE score.

associated with income and social status-like measures while it seems to be independent of measures of one's own individual characteristics, such as SEN, ethnicity and disability.

<<<

Table 3>>>

A further analysis of the relationships between these disadvantage indicators is given in Table 4 as a tetrachoric correlation matrix, which is preferable for binary outcomes (Brown 1977). The entries in the body of the table represent the correlation of experiencing the named socioeconomic disadvantage given the child is growing up in a family with one of the other disadvantages.

<<<Table 4 >>>

We find that there is substantial variation in the correlations linking the various Widening Participation indicators in the Next Steps sample. The highest correlation coefficient is found between income deprivation (IDACI) and multiple deprivation (IMD) (0.95), which is intuitive in the sense that both are postcode level measures of social background, and one is part of the other. We find moderate correlations between being Non-White and income deprivation (0.56); low SES and income deprivation (0.41) and multiple deprivation (0.44). Being a potential FiF is weakly correlated with low SES (0.34), multiple (0.31) and income deprivation (0.28), and FSM eligibility (0.29), and uncorrelated with the rest of the measures (non-significant correlation coefficient or significant correlation coefficient below 0.2). This is perhaps unsurprising given that potential FiF make up a large proportion of the sample and many of the other WP measures do not.

3. Empirical strategy

We assess potential FiF as a WP indicator by comparing how well it predicts HE participation and graduation compared to the aforementioned WP indicators. We examine potential FiF on its own and conditional on gender and prior educational attainment. In a just world, conditional on prior educational attainment, no WP measure should be significantly related neither to the probability of HE participation nor graduation. This does not necessarily imply that everybody would be better-off by going to university rather than choosing a different path, but it does mean that family background should not represent a systematic barrier to doing so. If WP measures do not create a systematic barrier, they should not be negative predictors of HE participation and graduation; however, we know (and show in this paper) that this is not the case and HE achievements are correlated with family background even on top of prior educational attainment. Our question is, therefore, how good of a predictor is being potential FiF in comparison to, and, on the top of the other commonly used WP measures. If it is a relatively 'good' predictor, it should be included in WP policies. If, however, this is not the case, i.e. the above listed WP measures capture all potential sources of disadvantage and the inclusion of potential FiF does not provide new or better information on the barriers to HE, its use would not provide any additional benefit.

There is a wide literature on how to assess the relative importance of predictors. If we wanted to predict continuous dependent variables using measures that are independent of each other, we could estimate linear regressions and straightforwardly compare the predictive power of the explanatory variables by comparing the R-squared estimates of models including and not-including potential FiF, or, we could use variance decomposition techniques to see which variable explains the highest share of the variance of the dependent variable (Grömping 2009). However, (1) we predict binary variables and (2) the explanatory variables are not just correlated, but as potential FiF is the only direct measure of parental human capital among the WP indicators, it is

likely that some WP indicators (income deprivation, FSM eligibility, etc.) could already be the consequence of parental graduation, introducing a *bad control* problem (Angrist and Pischke 2008) into the models. Having bad controls among the explanatory variables causes the estimated coefficients to be biased and thus uninterpretable in the traditional regression framework.

The first challenge is easy to solve using non-linear probability models. If we compare the predictive power of the indicators estimated in separate logit models, having one measure in the model at a time, we do not have to worry about the bad control problem. Thus, in the first part of the analysis, we estimate simple logit models to predict university participation and graduation and compare the predictive power of each WP indicator estimated in separate models using the Area Under the Curve (AUC) measure (see its description in the next subsection). As we are also interested in the predictive power of potential FiF over and above other WP measures, in the second part, we add potential FiF to an empirically optimal set of WP indicators in multiple logit models, similarly to Ilie, Sutherland, and Vignoles (2017), and examine whether it improves their predictive power using again the AUC measure. This approach gives an answer to the question of whether adding potential FiF on top of a set of other measures gives new information on the barriers of HE participation and graduation, but it does not allow us to rank the measures in terms of their importance in making a prediction due to having bad controls among the explanatory variables. Thus, in the third part of the analysis, we employ a non-parametric approach, a random forest classification algorithm, to rank the measures in terms of their importance. As we will discuss, the random forest ranks the predictors in terms of their importance even if they are not independent of each other (Grömping 2009). Lastly, as our results from all three of these methods suggest that earlier educational attainment could be the main channel of the effect of family background on HE success, in the fourth part of the analysis we look at the ranking of WP measures in terms of predicting early educational attainment at ages 11, 13 and 16 using again the random forest approach. Each method used in the analysis is explained in more detail in the following sections.

3.1 Comparing the predictive power of WP measures in logit models, one-by-one

We start by comparing the predictive power of WP measures one-by-one, estimating simple logit models where the outcome variable, either HE participation or graduation, is regressed on each WP indicator separately. Several pseudo-R^2-type measures could be used with logit models to investigate the fit of models, but all have their drawbacks (Estrella 1998). Thus, we compare the predictive power of each model using the AUC, which is the estimated Area under the Receiver Operating Characteristic (ROC) curve (Kuhn and Johnson 2013; p. 262-265). The AUC is a measure to compare how well models predict a binary outcome variable. The quality of prediction is captured by the share of the real binary outcome predicted correctly by the model; the higher the share, the better the model. The AUC summarises this information based on the share of correctly and incorrectly predicted cases when the outcome variable equals to 1 (share of true positives and share of false positives), along all potential threshold between 0 and 1 that could be used to convert the continuous values of the predicted (latent) variable to a binary prediction. The AUC can be estimated to assess predictive power with any models that predict a binary variable, so it is directly comparable across different specifications of logit and random forest classification models, which is especially useful for our purposes. It is also fairly easy to interpret: its value is between 0 and 1 and flipping a coin would produce an AUC of 0.5. As a rule of thumb, the predictive power of a model is considered good if AUC>0.8 and great if AUC>0.9.

As mentioned before, schools are the main sampling unit of Next Steps and students were chosen within schools. To control for potential heterogeneity across schools, in addition to a set of simple logit models we also estimate a set of logit models where we control for school characteristics; in particular, the sampling school averages of Key Stage 2 total scores and capped linear GCSE scores of an individual's peers. Peers' Key Stage 2 scores are measured at age 11 in primary schools, before students progressed to secondary school, and we interpret them in our framework as a proxy for sampling 'school intake', while GCSE scores, conditional on Key Stage 2 results, are interpreted as a proxy for the compulsory school progression of peers, i.e. 'school value added'. This is preferable to a random effects or fixed effects model due to the small number of observations per school (see Appendix B for further information). We believe that these measures capture school-level shocks that are related to student intake and school quality. However, there is still concern that even these school-level variables might be bad controls as school choice could be the result of parental education. If this is the case, we will estimate lower marginal predictive power of being potential FiF in the logit models with school-level averages than in the simple logit models because school characteristics would already absorb some of the effects of parental background. Thus, we interpret our results from the simple logit models as the *upper bound* while the results for the logit models with school-level controls as the *lower bound* of the marginal predictive power of potential FiF. As before, we use the quintiles of the continuous school averages of one's peers and include a sixth category for the missing values.

Additionally, we estimate a set of models both with and without controlling for individual-level early educational attainment and gender. Formally, we estimate logit models as:

$$f(outcome_i) = \propto_0 + \beta * indicator_i + \gamma * X_i + \delta * Z_{is}\,; \qquad (1)$$

where

$outcome_i$	is a binary variable capturing whether individual i ever went to university or graduated by age 25 (1), or not (0);
$\propto_0$	is the intercept;
$indicator_i$	is one of the WP indicator variables for individual i;
X_i	is a vector of individual characteristics for individual i (KS2 and KS4 scores, gender);
Z_{is}	is a vector of the school-level average of KS2 and KS4 scores of the peers of individual i in school s;
f()	is the logistic function.

In all models we estimate robust standard errors clustered at the level of the school and report odds ratios and 95% confidence intervals.

3.2 Looking at whether adding potential FiF on top of a parsimonious set of WP measures in logit models increases predictive power

In the second part of the analysis, we look at whether adding potential FiF on the top of an empirically optimal set of the most widely used WP measures increases the predictive power of models explaining the probability of HE participation and graduation. We choose an empirically

optimal parsimonious set of WP measures to set a baseline model using a Least Absolute Shrinkage and Selection Operator (Lasso) procedure (Friedman, Hastie, and Tibshirani 2009). As we have shown, WP measures are correlated with each other and have a substantial overlap. The Lasso procedure aims at finding the most parsimonious model by constraining some of the estimated coefficients of explanatory variables to zero. In particular, the Lasso penalises the magnitude of coefficients when optimising the maximum likelihood function of the logit model. We choose to work with a parsimonious set of explanatory variables to avoid model over-fitting due to having cross-correlated explanatory variables on the right-hand side. We apply the Lasso procedure to a logit model that includes all WP measures used before as explanatory variables on the right hand-side, except the potential FiF measure. Having the empirically optimal set of WP indicators, we add the potential FiF measure on top these in multiple logit models and look at whether it increases the predictive power of the model using the AUC measure introduced before.

3.3 Ranking the predictors of HE participation and graduation in terms of their importance

In the third part, we use a non-parametric algorithm, a random forest, to predict HE participation and graduation and look at the relative importance of WP measures in this prediction. The random forest classification algorithm works by constructing series of decision trees and predicting the outcome from each series as the modes of predictions (Breiman 2001). A decision tree is an algorithm that repeatedly splits the data to subsamples (*branches*) along certain values of the explanatory variables in order to create as homogenous clusters (*leaves*) in terms of the outcome variable as possible. The predicted outcomes of a decision tree depend heavily on the order of variables used to split the sample; thus, the random forest algorithm randomly chooses different explanatory variables to start with on bootstrapped samples of the data and appoints the most frequent predicted class from all repetitions.

The main advantage of using the random forest algorithm is that it provides a straightforward ranking of variables in terms of how important they are in predicting the outcome even if they are not independent. The ranking is produced based on the out-of-sample predictive performance of the measures obtained through k-fold cross-validation (i.e. averaged out over bootstrapped samples) and thus it is resistant to over-fitting. A variable is considered more important if splitting the sample based on the variable leads to more homogenous subgroups in terms of the outcome variable than splitting the sample based on another. The more homogenous, or in other words, the less heterogeneous the resulting subgroups are, the better the variable classifies the observations according to the categories of the outcome variable. Thus, we measure the relative importance of explanatory variables by the Mean Decrease in Gini measure, which captures how well the variable decreases the heterogeneity of subgroups by splitting the sample on a given variable averaged across all decision trees (Friedman, Hastie, and Tibshirani 2009). We construct 95% confidence intervals around the estimated Mean Decrease in Gini parameters via bootstrapping.

3.4 Ranking the predictors of early educational attainment in terms of their importance

Lastly, as our results suggest that early educational attainment might be the most important channel of the relationship between potential FiF (and all other WP measures) and HE participation and graduation, we directly investigate this hypothesis using random forest regression (as opposed to classification) models predicting age 16 (GCSE), age 14 (Key Stage 3) and age 11

(Key Stage 2) school exam scores and look at the relative importance of potential FiF in these predictions. As we now predict continuous outcome variables in these random forest models, and not binary outcomes as before, we use the % Increase in Mean Squared Errors (MSE) measure, which captures the increase in MSE should predictors be replaced by their own randomly permuted values (Friedman, Hastie, and Tibshirani 2009). We construct 95% confidence intervals around the estimated parameters via bootstrapping.

4. Results

4.1 Comparing the predictive power of WP measures in logit models, one-by-one

We begin by estimating a series of simple logit models and compute the AUC for each outcome. For each WP indicator and outcome, we run both a simple logit model and a logit model with school-level controls for student intake and school quality. As mentioned before, the simple logit models capture the *upper bound* while the logit models with school controls capture the *lower bound* of the marginal predictive power of WP measures.

<<<Tables 5 and 6>>>

Tables 5 and 6 show the estimated coefficients as odds ratios, along with their 95% confidence intervals and p-values from all four logit models. All odds ratios in Tables 5 and 6 are estimated in separate models. The odds ratios of potential FiF are the smallest in all four types of models for both outcomes. In terms of predicting the probability of HE participation, although the estimated odds ratio goes up from 0.275 to 0.507 as more controls are added to the model, it stays highly significant and produces the lowest p-value even in the fourth model (Table 5, Model 4). In terms of predicting the probability of HE graduation, the odds ratio of potential FiF is somewhat higher, between 0.436 and 0.670, but it is still the lowest in magnitude and has the lowest p-values among the WP measures (Table 6). This indicates that potential FiF is an important predictor of HE outcomes as compared to other WP measures.

<<<Figures 2 and 3>>>

Figures 2 and 3 show the predictive power of WP measures and potential FiF status, captured by the area under the ROC curve, estimated by the same probability models as detailed above and shown in Tables 5 and 6. As a point of comparison, a random binary variable was also created and included among the 'real' measures. In the first logit model, potential FiF has the highest relative predictive power in predicting both HE participation and graduation (AUC=0.59, Figure 2, left panel, and, AUC=0.56, Figure 2, right panel). Once sampling school-level averages of peers' KS2 and GCSE scores and school type are included in the models, the difference in the predictive power of measures decreases, but still potential FiF has the highest predictive power (AUC=0.65, Figure 2, left panel, and, AUC=0.60, Figure 2, right panel). Again, this indicates that potential FiF is a strong predictor as compared to other WP measures.

Figure 3 extends the same models by adding individual-level control variables. This includes pre-university educational attainment measures (KS2 quintiles, GCSE quintiles) and gender. Adding these control variables (especially pre-university educational attainment) increases the predictive power of the logit models to between AUC=0.74 and AUC=0.76 in the case of HE participation (Figure 3, left panel) and AUC=0.66 in the case of HE graduation (Figure 3, right panel). After controlling for early educational attainment, the predictive power of the models does not change even with the inclusion of WP measures. This is in spite of the fact that some of the

measures have significant coefficients in the models (see Table 6). Thus, the significance of the coefficients in the model on its own does not say anything about their predictive power. These results also show that once prior educational attainment is taken into account, none of the WP measures is significantly better on their own than a random binary variable at predicting university participation and graduation.

4.2 Looking at whether adding potential FiF on top of a parsimonious set of WP measures in logit models increases predictive power

In the second part of the analysis, we look at whether adding potential FiF to an empirically optimal combination of WP measures increases the predictive power of the models explaining the probability of HE participation and graduation. We choose the baseline model to be extended by potential FiF using a Lasso procedure (see Table A2 in the Appendix for further information). In terms of predicting HE participation, SEN, FSM, low SES, income deprivation, disability, single HH, and being a care leaver are selected to be important, while in terms of predicting the probability of HE graduation, SEN, FSM, low SES, single HH and being a carer leaver remain in the parsimonious model.[12]

<<<Tables 7 and 8>>>

The results in Tables 7 and 8 show that the potential FiF measure has a relatively large, highly significant negative association with HE outcomes in all extended models. Having parents who are not university graduates decreases the relative odds of HE participation to 0.547 (Table 7, last column) and the relative odds of HE graduation to 0.703 (Table 8, last column) over and above controlling for a large set of WP measures, individual early educational attainment and a proxy for secondary school quality. These results suggest that potential FiF might capture additional sources of disadvantage beyond other measures, although due to the bad control problem that we discussed above, the interpretation of the coefficients estimated from the multiple models is limited.

<<<Figures 4 and 5>>>

We assess the marginal predictive power of adding potential FiF to the parsimonious model using the AUC measure in Figures 4 and 5. Extending the baseline models with potential FiF increases their predictive power both in terms of predicting the probability of HE participation and graduation. Adding potential FiF to the model predicting HE participation increases the AUC from 0.65-0.68 to 0.68-0.70 and from 0.60-0.62 to 0.62-0.63 when predicting HE graduation if we do not control for early educational attainment. These improvements are significantly greater than zero in the models that do not control for school-level variables (Figure 4, first vs. second bar on the left and right panel), but they are not significant in the models that include school-level variables (Figure 4, third vs. fourth bar on the left and right panel). Controlling for early educational attainment decreases the marginal predictive power of adding potential FiF to the model to close to zero, in spite of the fact that the actual estimated coefficients on potential FiF stays significant and negative (Tables 7 and 8). We show that these results are robust to the type of modelling strategy chosen in Appendix D.

[12] Note that if FiF is also included in the baseline model, the Lasso procedure would not push its coefficient to zero.

4.3 Ranking the predictors of HE participation and graduation in terms of their importance

Next, we turn to a non-parametric empirical strategy, a random forest classification algorithm, to rank the predictors of HE participation and graduation in terms of their importance in making these predictions. As measure of variable importance, we use the Mean Decrease in Gini, which captures how well the variables classify the observations on average across all decision trees of the random forest (Friedman, Hastie, and Tibshirani 2009).

<<<Figures 6 and 7>>>

Figures 6 and 7 compare the relative importance of measures (including potential FiF) without (left panel) and with controlling for early educational attainment (right panel.) Figure 6 shows that without controlling for early educational attainment, the potential FiF measure ranks first in terms of its importance captured by the Mean Decrease in Gini to predict HE participation (left panel). When early educational attainment is added to the model, the importance of GCSE and KS2 scores turn out to be highest, and the importance of all other WP measures in the model is of about the same magnitude (Figure 6, right panel). Similarly, potential FiF ranks first in terms of its importance in predicting HE graduation (Figure 7, left panel), but when early educational attainment is added to the model the relative importance of all WP measures becomes very similar and much lower than the importance of early educational attainment.

4.4 Ranking the predictors of early educational attainment in terms of their importance

We have seen that while potential FiF has the highest predictive power in both a logit and a random forest framework, once early educational attainment is included, its additional predictive power (as well as the predictive power of all other WP measures) diminishes. As has been mentioned before, early educational attainment already captures some effects of parental background and our analysis suggests that indeed being a potential FiF affects early educational attainment. To investigate this, we look at the relationship of WP measures and early educational attainment using random forest regressions to predict school exam scores captured at age 16 (standardised capped linear GCSE score), age 14 (total Key Stage 3 score) and age 11 (total Key Stage 2 score). As we now have continuous dependent variables in these models, we apply random forest regression (as opposed to classification) models and use a different measure of predictive power, the percentage increase in the Mean Squared Errors (MSE). This measure is estimated out-of-sample through a cross-validation procedure.

We estimate five random forest regression models:

1. we predict age 16 (GCSE)scores using WP measures;
2. we predict age 16 (GCSE) scores using WP measures and controlling for age 14 (KS3) scores;
3. we predict age 14 (KS3) scores using WP measures;
4. we predict age 14 (KS3) scores using WP measures and controlling for age 11 (KS2) scores;
5. we predict age 11 (KS2) scores using WP measures.

<<<Figures 8, 9, and 10>>>

Figures 8, 9, and 10 show the importance of WP measures in predicting national exam scores without and with controlling for the previous early educational attainment measures. In terms of predicting age 16 (GCSE) scores, potential FiF turns out to be the most important with percent increase in MSE= 99.22 (Figure 8, left panel), but once earlier attainment is introduced, this earlier attainment measure becomes the most important (percent increase in MSE= 126.87) and the relative importance of all WP measures decreases drastically (Figure 8, right panel). Similarly, potential FiF and SEN turn out to be relatively more important than the rest of the measures in predicting age 14 (KS 3) scores (Figure 9, left panel), but again once earlier attainment is introduced, it becomes the most important predictor of the model (Figure 9, right panel). Lastly, in predicting age 11 (KS2) scores, SEN and potential FiF are the most important and relatively far more important than all other measures (Figure 10). In Appendix D we show as an additional robustness check that this result is not dependent on the modelling strategy used. Unfortunately, our data do not allow us to control for earlier test scores or any measures of cognitive abilities. This final result provides some suggestive evidence that potential FiF is an important predictor of pre-university attainment.

5. Discussion and conclusion

The analysis presented in this paper is the first step in unpacking the extent to which using 'first in family' as an indicator captures the same or different individuals as the other sociodemographic characteristics used by universities in their Widening Participation agendas. The fact that universities have been using potential FiF as a Widening Participation indicator without any exploration of its validity as an indicator has prompted this paper.

Universities already make use of information on prospective students' socioeconomic background, the types of school they attended and their national exam results in order to inform their admission process. Our results show that being a potential FiF overlaps more with family-background and income-type measures of disadvantage (income deprivation, living in a single household, low SES) while it is independent from some individual-level characteristics as SEN or disability. In a logit modelling framework, we find that the potential FiF measure is certainly not worse (if anything, better) than other measures of disadvantage like FSM eligibility, low social class or living in a single household, in predicting HE participation or graduation. In fact, adding potential FiF to an empirically optimal combination of WP measures increases the predictive power of a model. Using random forest models to predict HE participation and graduation reveals that the potential FiF measure is the most important predictor of HE success.

Once we control for early educational attainment in our models, however, the additional predictive power of potential FiF (as well as any other measure) becomes negligible. It performs no better or worse than any other WP indicators. Thus, we hypothesise that being potential FiF affects HE outcomes mainly through early educational attainment. Indeed, when we look at the predictive power of WP measures on pre-university educational attainment, being potential FiF and having special educational needs (SEN) are the most important predictors. We believe that our finding on early educational attainment being the main channel of the effect of social disadvantage on HE success might supply evidence to support the need for Contextual Admissions using individual level measures of disadvantage.

It is not surprising that being potential FiF seems to be an important barrier to HE participation and graduation and a key determinant of early educational attainment. If we assume that education increases human capital, being a potential FiF, i.e. the fact that one's parents have not graduated from university, can be viewed as a measure of parental human capital, while the other measures of disadvantaged family background, i.e. low social class, income deprivation, or

FSM, which proxy for family income, could very well already be the consequence of parental education. It is also interesting that potential FiF and SEN, both individual level characteristics, end up being more important than area level measures of disadvantage (e.g. IDACI and IMD), especially in light of the current debate surrounding area-level measures such as POLAR.

We acknowledge that there should be some consideration of the nature of the WP measures from a practical point of view, for example, taking into account how 'gameable' the measures are. This reopens the debate about whether they are verifiable, accurate, and reliable. Although HE participation is highly dependent on the individual achievements of students, their social background and the current policies of universities also matter. When we predict the probability of HE participation, for example, we predict the joint probability that one applies to a university and gets accepted. The fact that after controlling for early educational attainment disadvantage measures do not affect the predictive power of the models predicting HE participation might mean both that 1) early educational attainment determines HE participation on its own as it already captures the effects of social background, and 2) as universities do apply Widening Participation policies, in practice, the potentially positive effects of these policies counteract the negative effects of social background once we control for early educational attainment. The linking of administrative (UCAS) data on university application and participation separately could help in disentangling these two channels.

Our analysis contributes to the applied empirical literature on education policy by looking at the relevance of the potential FiF measure using a range of quantitative methods and paying special attention to the data we use. We draw two main conclusions. First, the potential FiF indicator seems to be just as valid a WP measure as other measures, and it is informative above and beyond the usually used indicators as well. Potential FiF is an important barrier to university participation and graduation, even after controlling for other sources of disadvantage. This seems to work through the channel of early educational attainment. Second, our research also provides evidence that the potential FiF indicator could be key in efforts to widen participation at universities through the use of contextualised admissions; however, the predictive power of all WP measures altogether is surprisingly low. A large share of the individual heterogeneity of HE success is still unexplained, even after controlling for early educational attainment, and thus should be the subject of further research.

References

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

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Belfield, Chris, Jack Britton, Franz Buscha, Lorraine Dearden, Matt Dickson, Laura van der Erve, Luke Sibieta, Anna Vignoles, Ian Walker, and Yu Zhu. 2018. 'The Impact of Undergraduate Degrees on Early-Career Earnings'. 27 November 2018. https://www.ifs.org.uk/publications/13731.

Blanden, Jo, and Stephen Machin. 2004. "Educational Inequality and the Expansion of UK Higher Education." *Scottish Journal of Political Economy* 51 (2): 230–49.

Blanden, Jo, and Lindsey Macmillan. 2016. "Educational Inequality, Educational Expansion and Intergenerational Mobility." *Journal of Social Policy* 45 (4): 589–614. https://doi.org/10.1017/S004727941600026X.

Boliver, Vikki. 2013. "How Fair Is Access to More Prestigious UK Universities?" *The British Journal of Sociology* 64 (2): 344–64. https://doi.org/10.1111/1468-4446.12021.

Boliver, Vikki, C. Crawford, M. Powell, and W. Craige. 2017. "Admissions in Context: The Use of Contextual Information by Leading Universities." The Sutton Trust. https://www.suttontrust.com/wp-content/uploads/2017/10/Admissions-in-Context-Final_V2.pdf.

Boliver, Vikki, Stephen Gorard, Nadia Siddiqui, Vikki Boliver, Stephen Gorard, and Nadia Siddiqui. 2015. "Will the Use of Contextual Indicators Make UK Higher Education Admissions Fairer?" *Education Sciences* 5 (4): 306–22. https://doi.org/10.3390/educsci5040306.

Bowen, William G., Matthew M. Chingos, and Michael S. McPherson. 2009. *Crossing the Finish Line: Completing College at America's Public Universities*. Princeton University Press.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Britton, Jack, Lorraine Dearden, Neil Shephard, and Anna Vignoles. 2016. "How English Domiciled Graduate Earnings Vary with Gender, Institution Attended, Subject and Socio-Economic Background." April 13, 2016. https://www.ifs.org.uk/publications/8233.

Brown, Morton B. 1977. "Algorithm AS 116: The Tetrachoric Correlation and Its Asymptotic Standard Error." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 26 (3): 343–51. https://doi.org/10.2307/2346985.

Centre for Longitudinal Studies. 2018. "Next Steps Age 25 Survey: User Guide." Centre for Longitudinal Studies. https://cls.ucl.ac.uk/wp-content/uploads/2017/11/5545age_25_survey_user_guide-1.pdf.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, Emmanuel Saez, and Nicholas Turner. 2014. "Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility." *American Economic Review* 104 (5): 141–47. https://doi.org/10.1257/aer.104.5.141.

Chowdry, Haroon, Claire Crawford, Lorraine Dearden, Alissa Goodman, and Anna Vignoles. 2013. "Widening Participation in Higher Education: Analysis Using Linked Administrative Data: *Widening Participation in Higher Education*." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176 (2): 431–57. https://doi.org/10.1111/j.1467-985X.2012.01043.x.

Clarke, Paul, Claire Crawford, Fiona Steele, and Anna Vignoles. 2015. "Revisiting Fixed- and Random-Effects Models: Some Considerations for Policy-Relevant Education Research." *Education Economics* 23 (3): 259–77. https://doi.org/10.1080/09645292.2013.855705.

Crawford, Claire. 2014. "Socio-Economic Differences in University Outcomes in the UK: Drop-out, Degree Completion and Degree Class." IFS. https://doi.org/10.1920/wp.ifs.2014.1431.

Currie, Janet, and Enrico Moretti. 2003. "Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings." *The Quarterly Journal of Economics* 118 (4): 1495–1532.

Department for Communities and Local Government. 2015. "The English Indices of Deprivation 2015." Department for Communities and Local Government. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/465791/English_Indices_of_Deprivation_2015_-_Statistical_Release.pdf.

Department for Education. 2011. "LSYPE User Guide to the Datasets: Wave 1 to Wave 7." http://doc.ukdataservice.ac.uk/doc/5545/mrdoc/pdf/5545lsype_user_guide_wave_1_to_wave_7.pdf.

———. 2017. "Statistics: Participation Rates in Higher Education." Participation Rates in Higher Education for England. September 28, 2017. https://www.gov.uk/government/collections/statistics-on-higher-education-initial-participation-rates.

———. n.d. "What Qualification Levels Mean." GOV.UK. Accessed June 28, 2019. https://www.gov.uk/what-different-qualification-levels-mean/list-of-qualification-levels.

Estrella, Arturo. 1998. "A New Measure of Fit for Equations With Dichotomous Dependent Variables." *Journal of Business & Economic Statistics* 16 (2): 198–205. https://doi.org/10.1080/07350015.1998.10524753.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second edition. Springer.

Gorard, Stephen. 2012. "Who Is Eligible for Free School Meals? Characterising Free School Meals as a Measure of Disadvantage in England." *British Educational Research Journal* 38 (6): 1003–17. https://doi.org/10.1080/01411926.2011.608118.

Gorard, Stephen, Vikki Boliver, Nadia Siddiqui, and Pallavi Banerjee. 2017. "Which Are the Most Suitable Contextual Indicators for Use in Widening Participation to HE?" *Research Papers in Education* 0 (0): 1–31. https://doi.org/10.1080/02671522.2017.1402083.

Gorard, Stephen, Nadia Siddiqui, and Vikki Boliver. 2017. "An Analysis of School-Based Contextual Indicators for Possible Use in Widening Participation." *Higher Education Studies* 7 (2): 79. https://doi.org/10.5539/hes.v7n2p79.

Grömping, Ulrike. 2009. "Variable Importance Assessment in Regression: Linear Regression versus Random Forest." *The American Statistician* 63 (4): 308–19. https://doi.org/10.1198/tast.2009.08199.

Henderson, Morag, Nikki Shure, and Anna Adamecz-Volgyi. 2019. "'First in the Family' University Graduates in England." *IZA Discussion Papers* No. 12588.

Hobbs, Graham, and Anna Vignoles. 2010. "Is Children's Free School Meal 'Eligibility' a Good Proxy for Family Income?" *British Educational Research Journal* 36 (4): 673–90. https://doi.org/10.1080/01411920903083111.

Holmlund, Helena, Mikael Lindahl, and Erik Plug. 2011. "The Causal Effect of Parents' Schooling on Children's Schooling: A Comparison of Estimation Methods." *Journal of Economic Literature* 49 (3): 615–51. https://doi.org/10.1257/jel.49.3.615.

Ilie, Sonia, Alex Sutherland, and Anna Vignoles. 2017. "Revisiting Free School Meal Eligibility as a Proxy for Pupil Socio-Economic Deprivation." *British Educational Research Journal* 43 (2): 253–74. https://doi.org/10.1002/berj.3260.

Kuhn, Max, and Kjell Johnson. 2013. *Applied Predictive Modeling*. Springer Science & Business Media.

Maurin, Eric, and Sandra McNally. 2008. "Vive La Revolution! LongTerm Educational Returns of 1968 to the Angry Students." *Journal of Labor Economics* 26 (1): 1–33. https://doi.org/10.1086/522071.

Office for National Statistics. n.d. "The National Statistics Socio-Economic Classification (NS-SEC) - Office for National Statistics." The National Statistics Socio-Economic Classification (NS-SEC). Accessed July 11, 2019. https://www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/thenationalstatisticssocioeconomicclassificationnssecrebasedonsoc2010.

Office For National Statistics, Social Survey Division. 2019. "Annual Population Survey, July 2015 - June 2016." UK Data Service. https://doi.org/10.5255/UKDA-SN-8054-5.

Oreopoulos, Philip, and Uros Petronijevic. 2013. "Making College Worth It: A Review of Research on the Returns to Higher Education." Working Paper 19053. National Bureau of Economic Research. https://doi.org/10.3386/w19053.

Rose, David, and David Pevalin. 2001. "The National Statistics Socio-Economic Classification: Unifyingofficial and Sociological Approaches to the Conceptualisation Andmeasurement of Social Class." *ISER Working Paper Series*, no. 2001–04.

Selby, John. 2018. "HEFCE History: The Early Days of Widening Participation." *WONKHE* (blog). May 3, 2018. https://wonkhe.com/blogs/hefce-history-the-birth-of-widening-participation/.

Suhonen, Tuomo, and Hannu Karhunen. 2017. "The Intergenerational Effects of Parental Higher Education: Evidence from Changes in University Accessibility." *VATT Working Paper*, no. 100 (December). http://www.doria.fi/handle/10024/148937.

Vignoles, Anna F, and Nattavudh Powdthavee. 2009. "The Socioeconomic Gap in University Dropouts." *The B.E. Journal of Economic Analysis & Policy* 9 (1). https://doi.org/10.2202/1935-1682.2051.

Walker, Ian, and Yu Zhu. 2018. "University Selectivity and the Relative Returns to Higher Education: Evidence from the UK." *Labour Economics* 53 (August): 230–49. https://doi.org/10.1016/j.labeco.2018.05.005.

Wyness, Gill. 2010. "Policy Changes in UK Higher Education Funding, 1963-2009." 10–15. DoQSS Working Papers. Department of Quantitative Social Science - UCL Institute of Education, University College London. https://ideas.repec.org/p/qss/dqsswp/1015.html.

Table 1. Comparison of HE participation and graduation rates of those aged 25/26 in 2015 in Next Steps vs. other sources

	HE participation	HE graduation rate
Next Steps[#], unweighted	51.7%	34.9%
Next Steps[#], weighted by wave-8 weights	40.0%	26.9%
Higher Education Initial Participation Rate (HEIPR), age 17-25*	40.5%	-
HE graduation rate, 2015 APS**	-	39.6%

HE participation: the share of those ever been to university as a % of the total population (aged 25/26 in England).
Higher Education Initial Participation Rate (HEIPR): the estimated likelihood of a young person participating in HE by age 25/26, i.e. the sum of the likelihoods of HE entry at ages 17-25/26, in England.
HE graduation rate in APS: the share of those having a Level-4 degree among those aged 25/26 in England.
HE graduation rate in Next Steps: the share of those having a university degree (BA/BSc or higher degrees).
Sources: #University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016 *Department for Education (2017) **Annual Population Survey, 2015. England only, those aged 25/26, weighted. (Office For National Statistics 2019).

Table 2. The prevalence of socioeconomic disadvantage used by the Widening Participation agenda among those having any Level 3 education

Disadvantage (age of measurement)	No. of observations with non-missing data on the disadvantage	Proportion of young people with non-missing data facing the disadvantage	Proportion of these who face at least one more disadvantage
Potential FiF (by age 17)	5,020	75.8%	81.6%
Single household (by age 17/18)	5,047	41.5%	92.4%
FSM (age 13/14-17/18)	4,270	10.0%	99.8%
Low social class (age 13/14-17/18)	4,971	40.2%	94.9%
Young carer (age 13/14-15/16)	4,931	9.6%	96.7%
Non-White	5,047	19.0%	96.2%
SEN (age 13/14-16/17))	5,023	16.4%	91.7%
Disabled (age 13/14-16/17)	5,025	5.7%	96.2%
Care leaver (age 13/14-17/18)	4,535	5.2%	98.8%
Multiple deprivation (age 14/15-15/16)	4,666	14.6%	99.9%
Income deprivation (age 14/15-15/16)	4,666	13.8%	100.0%

Notes: Weighted using Wave 8 final weights. N= 5,047. Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

Table 3. The overlap between potential FiF status and other WP measures: the share of potential FiF among those facing and not facing disadvantage measures

Disadvantage	Share of potential FiF among those facing the disadvantage	Share of potential FiF among those not facing the disadvantage	Diff, pp	t-test p-value
Single household	83.3%	73.5%	9.8	0.000
FSM	92.1%	76.3%	15.8	0.000
Low social class	86.7%	71.4%	15.3	0.000
Young carer	84.1%	76.9%	7.2	0.001
Non-White	77.7%	77.6%	0.1	0.966
SEN	77.8%	77.6%	0.1	0.944
Disabled	75.2%	77.8%	-2.5	0.464
Care leaver	87.0%	77.1%	9.8	0.000
Multiple deprivation	90.5%	75.6%	14.9	0.000
Income deprivation	90.3%	75.8%	14.5	0.000

Notes: Complete cases only. N=3,880. Weighted using Wave 8 final weights. Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

Table 4. Tetrachoric correlations of disadvantage

Variables	1	2	3	4	5	6	7	8	9	10	11
(1) FIF	1	0.15***									
(2) Single HH	0.15***	1	0.39***								
(3) FSM	0.29***	0.39***	1								
(4) Low social class	0.34***	-0.01	0.7***	1							
(5) Young carer	0.14*	0.12	0.28***	0.16***	1						
(6) Non-white	0.07	0.04	0.47***	0.32***	0.18***	1					
(7) SEN	-0.02	0.05	0.03	-0.01	0.02	-0.19***	1				
(8) Disabled	-0.04	0.1	0.14	0.01	0.09	-0.05	0.5***	1			
(9) Care leaver	0.13	0.36***	0.27***	0.14	0.08	0.21***	0.02	0.15	1		
(10) Multiple deprivation	0.31***	0.23***	0.56***	0.44***	0.08	0.55***	-0.08	0.02	0.22***	1	
(11) Income deprivation	0.28***	0.3***	0.6***	0.41***	0.09	0.56***	-0.07	0.02	0.21**	0.95***	1

Notes: Complete cases only. Unweighted. N=3,880. Statistical significance is tested while correcting for the number of hypotheses tested together using the Bonferroni correction. *p<0.05, **p<0.01, ***p<0.001. Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

Table 5. The relationship between HE participation and WP measures (simple logit models)

	Logit models		Logit models with school-average controls		Logit models with individual-level controls		Logit models with school-average and individual-level controls	
	Model 1		Model 2		Model 3		Model 4	
	Odds ratio	p-value	Odds ratio	p-value	Odds ratio	p-value	Odds ratio	p-value
Potential FiF	0.275*** [0.110; 0.690]	0.006	0.353*** [0.210; 0.594]	0.000	0.463*** [0.315; 0.680]	0.000	0.507*** [0.361; 0.712]	0.000
Single HH	0.583*** [0.311; 1.091]	0.091	0.619*** [0.487; 0.787]	0.000	0.719*** [0.583; 0.886]	0.002	0.733*** [0.610; 0.882]	0.001
FSM	0.595* [0.348; 1.016]	0.057	0.719** [0.542; 0.954]	0.022	1.062 [0.795; 1.419]	0.698	1.127 [0.835; 1.523]	0.442
Low social class	0.651** [0.440; 0.962]	0.031	0.733*** [0.610; 0.882]	0.001	0.852* [0.712; 1.020]	0.08	0.878 [0.738; 1.045]	0.144
Young carer	0.779 [0.544; 1.114]	0.172	0.869 [0.684; 1.105]	0.256	1.000 [1.000; 1.000]	0.987	1.020 [0.775; 1.342]	0.895
Non-White	1.733 [0.896; 3.352]	0.102	2.138*** [1.463; 3.126]	0.000	2.363*** [1.538; 3.632]	0	2.801*** [1.674; 4.687]	0.000
SEN	0.458** [0.243; 0.865]	0.016	0.463*** [0.315; 0.680]	0.000	0.844 [0.675; 1.055]	0.136	0.811** [0.648; 1.013]	0.065
Disabled	0.482** [0.235; 0.988]	0.046	0.554*** [0.382; 0.805]	0.002	0.787 [0.542; 1.141]	0.208	0.819 [0.565; 1.186]	0.294
Care leaver	0.527** [0.264; 1.055]	0.07	0.607** [0.404; 0.911]	0.016	0.684 [0.434; 1.077]	0.101	0.733 [0.472; 1.140]	0.169
Multiple deprivation	0.600** [0.363; 0.994]	0.047	0.819 [0.642; 1.044]	0.106	0.896 [0.692; 1.160]	0.411	1.020 [0.827; 1.258]	0.862
Income deprivation	0.600** [0.361; 0.999]	0.049	0.811* [0.637; 1.032]	0.088	0.896 [0.704; 1.140]	0.377	1.010 [0.830; 1.229]	0.927
Further control variables								
Individual early educational attainment	No		No		Yes		Yes	
Gender	No		No		Yes		Yes	
School controls	No		Yes		No		Yes	

No. of observations: 5,047. All coefficients are estimated in separate logit models to predict the probability of HE participation, controlling for missing flags. P-values are estimated based on robust standard errors clustered by sampling schools. 95% confidence intervals are in brackets. Coefficients are odds ratios. *** p<0.01, ** p<0.05, * p<0.1. Individual early educational attainment: capped linear Key stage 4 (GCSE) score quintiles and total Key stage 2 score quintiles. School controls: the average KS2 and KS4 test scores of individuals' peers in the sampling secondary schools, secondary school type. Weighted using final Wave 8 weights. Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

Table 6. The relationship between HE graduation and WP measures (simple logit models)

	Logit models		Logit models with school-average controls		Logit models with individual-level controls		Logit models with school-average and individual-level controls	
	Model 1		Model 2		Model 3		Model 4	
	Odds ratio	p-value	Odds ratio	p-value	Odds ratio	p-value	Odds ratio	p-value
Potential FiF	0.436*** [0.239; 0.797]	0.007	0.517*** [0.372; 0.719]	0	0.644*** [0.517; 0.802]	0.001	0.670*** [0.549; 0.819]	0
Single HH	0.726 [0.479; 1.101]	0.132	0.763*** [0.650; 0.896]	0.001	0.852* [0.725; 1.001]	0.051	0.861** [0.745; 0.994]	0.041
FSM	0.644* [0.399; 1.039]	0.071	0.741** [0.571; 0.962]	0.024	0.990 [0.767; 1.277]	0.944	1.010 [0.821; 1.243]	0.931
Low social class	0.684** [0.484; 0.966]	0.031	0.741*** [0.638; 0.861]	0	0.861* [0.740; 1.002]	0.052	0.869* [0.752; 1.004]	0.057
Young carer	0.756 [0.529; 1.080]	0.124	0.811 [0.648; 1.013]	0.065	0.932 [0.739; 1.176]	0.566	0.932 [0.751; 1.158]	0.537
Non-White	1.419 [0.892; 2.258]	0.14	1.553*** [1.246; 1.935]	0	1.682*** [1.297; 2.181]	0	1.804*** [1.343; 2.423]	0
SEN	0.583** [0.358; 0.950]	0.03	0.595*** [0.458; 0.771]	0	0.896 [0.716; 1.121]	0.341	0.887 [0.715; 1.100]	0.278
Disabled	0.691 [0.421; 1.133]	0.143	0.771 [0.560; 1.061]	0.11	1.020 [0.736; 1.414]	0.912	1.020 [0.790; 1.317]	0.887
Care leaver	0.589* [0.317; 1.093]	0.093	0.664** [0.454; 0.970]	0.034	0.756 [0.515; 1.110]	0.154	0.779 [0.528; 1.148]	0.208
Multiple deprivation	0.771 [0.550; 1.080]	0.131	0.951 [0.762; 1.187]	0.671	1.073 [0.856; 1.344]	0.555	1.127 [0.899; 1.414]	0.303
Income deprivation	0.733 [0.504; 1.067]	0.105	0.896 [0.717; 1.119]	0.338	1.000 [0.867; 1.154]	0.999	1.051 [0.844; 1.310]	0.669
Further control variables								
Individual early educational attainment	No		No		Yes		Yes	
Gender	No		No		Yes		Yes	
School controls	No		Yes		No		Yes	

No. of observations: 5,047. All coefficients are estimated in separate logit models to predict the probability of HE graduation, controlling for missing flags. P-values are estimated based on robust standard errors clustered by sampling schools. 95% confidence intervals are in brackets. Coefficients are odds ratios. *** p<0.01, ** p<0.05, * p<0.1. Individual early educational attainment: capped linear KS4 (GCSE) score quintiles and total Key stage 2 score quintiles. School controls: the average KS2 and KS4 test scores of individuals' peers in the sampling secondary schools, secondary school type. Weighted using final Wave 8 weights. Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

Table 7. Expanding the parsimonious model with potential FiF – HE participation (multiple logit models, odds ratios)

	Logit models		Logit models with school-average controls		Logit models with individual-level controls		Logit models with school-average and individual-level controls	
	Baseline	Extended	Baseline	Extended	Baseline	Extended	Baseline	Extended
SEN	0.509***	0.489***	0.516***	0.498***	0.948	0.901	0.931	0.896
	[0.416;	[0.397;	[0.421;	[0.403;	[0.753;	[0.712;	[0.739;	[0.708;
	0.623]	0.602]	0.632]	0.614]	1.195]	1.140]	1.174]	1.133]
FSM	0.881	0.901	0.878	0.895	1.164	1.157	1.106	1.107
	[0.651;	[0.666;	[0.644;	[0.659;	[0.831;	[0.831;	[0.791;	[0.795;
	1.191]	1.219]	1.197]	1.215]	1.631]	1.612]	1.547]	1.542]
Low social class	0.668***	0.763***	0.694***	0.775***	0.761***	0.826**	0.778***	0.833*
	[0.565;	[0.642;	[0.585;	[0.651;	[0.633;	[0.685;	[0.647;	[0.690;
	0.791]	0.906]	0.823]	0.923]	0.915]	0.996]	0.936]	1.005]
Income deprivation	0.649***	0.711**	0.803*	0.841	0.757**	0.788*	0.826	0.847
	[0.498;	[0.548;	[0.619;	[0.651;	[0.578;	[0.604;	[0.633;	[0.650;
	0.846]	0.923]	1.043]	1.087]	0.990]	1.029]	1.078]	1.104]
Non-White	2.368***	2.305***	2.625***	2.545***	2.821***	2.729***	3.146***	3.031***
	[1.916;	[1.861;	[2.116;	[2.043;	[2.238;	[2.166;	[2.481;	[2.391;
	2.926]	2.854]	3.256]	3.169]	3.555]	3.439]	3.987]	3.843]
Disabled	0.695**	0.691**	0.800	0.775	0.885	0.864	0.955	0.929
	[0.505;	[0.498;	[0.576;	[0.558;	[0.621;	[0.606;	[0.659;	[0.642;
	0.956]	0.959]	1.112]	1.077]	1.262]	1.232]	1.383]	1.346]
Single HH	0.632***	0.680***	0.660***	0.694***	0.731***	0.757***	0.751***	0.770***
	[0.542;	[0.580;	[0.565;	[0.591;	[0.615;	[0.636;	[0.633;	[0.648;
	0.737]	0.797]	0.772]	0.814]	0.867]	0.900]	0.891]	0.915]
Care leaver	0.584***	0.621**	0.622**	0.653**	0.634**	0.657*	0.678*	0.696
	[0.406;	[0.430;	[0.416;	[0.439;	[0.411;	[0.429;	[0.427;	[0.441;
	0.842]	0.898]	0.929]	0.972]	0.978]	1.007]	1.076]	1.096]
Potential FiF		0.316***		0.380***		0.493***		0.547***
		[0.258;		[0.308;		[0.397;		[0.439;
		0.387]		0.470]		0.611]		0.681]
Intercept	3.043***	7.099***	0.984***	2.347***	0.387***	0.728***	0.112***	0.211***
	[2.669;	[5.790;	[0.439;	[1.078;	[0.265;	[0.476;	[0.045;	[0.087;
	3.470]	8.704]	2.207]	5.110]	0.565]	1.114]	0.278]	0.514]
Further control variables								
Individual early educational attainment	No		No		Yes		Yes	
Gender	No		No		Yes		Yes	
School-level average Key stage 2 scores	No		Yes		No		Yes	

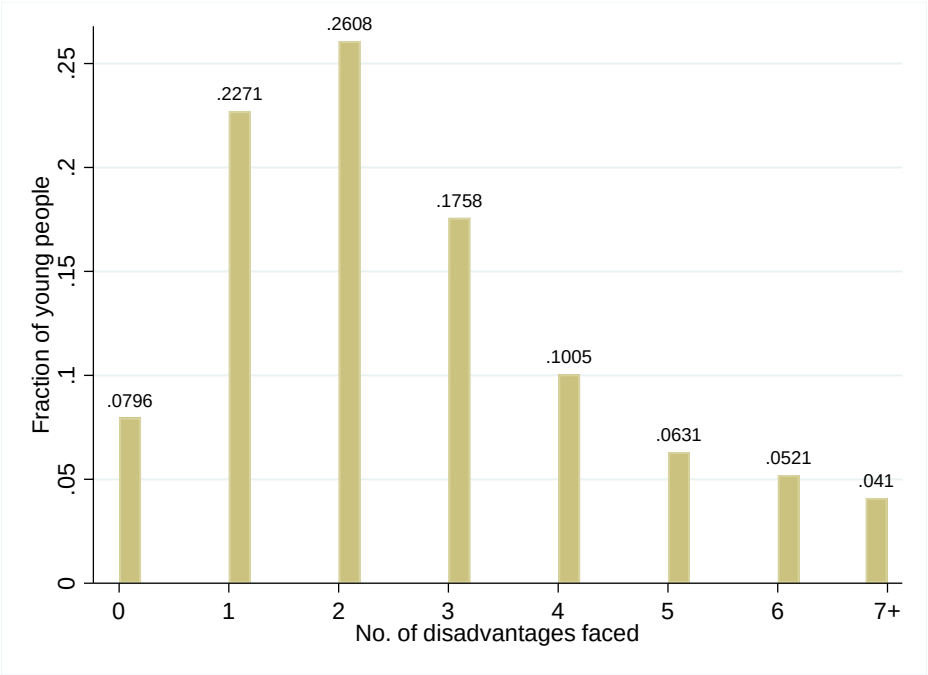
No. of observations: 5,047. The coefficients in each column are estimated in separate logit models to predict the probability of HE participation, controlling for missing flags. 95% confidence intervals constructed based on robust standard errors clustered by sampling schools are in brackets. Coefficients are odds ratios. *** p<0.01, ** p<0.05, * p<0.1. Individual early educational attainment: capped linear KS4 (GCSE) score quintiles and total Key stage 2 score quintiles. School controls: the average KS2 and KS4 test scores of individuals' peers in the sampling secondary schools, secondary school type. Weighted using final Wave 8 weights. Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

Table 8. Expanding the parsimonious model with potential FiF – HE graduation (multiple logit models, odds ratios)

	Logit models		Logit models with school-average controls		Logit models with individual-level controls		Logit models with school-average and individual-level controls	
	Baseline	Extended	Baseline	Extended	Baseline	Extended	Baseline	Extended
SEN	0.605***	0.593***	0.628***	0.613***	0.958	0.924	0.952	0.921
	[0.500;	[0.488;	[0.517;	[0.504;	[0.771;	[0.742;	[0.765;	[0.738;
	0.734]	0.720]	0.762]	0.746]	1.191]	1.151]	1.186]	1.150]
FSM	0.803	0.833	0.850	0.874	1.011	1.019	0.999	1.008
	[0.615;	[0.638;	[0.649;	[0.667;	[0.755;	[0.763;	[0.746;	[0.754;
	1.049]	1.087]	1.115]	1.145]	1.354]	1.362]	1.338]	1.347]
Low social class	0.715***	0.783***	0.733***	0.792***	0.799***	0.839**	0.803***	0.837**
	[0.618;	[0.675;	[0.633;	[0.683;	[0.688;	[0.722;	[0.691;	[0.720;
	0.829]	0.909]	0.850]	0.919]	0.927]	0.976]	0.931]	0.973]
Non-White	1.655***	1.642***	1.732***	1.697***	1.863***	1.842***	1.927***	1.893***
	[1.396;	[1.385;	[1.454;	[1.426;	[1.555;	[1.540;	[1.601;	[1.573;
	1.963]	1.947]	2.062]	2.021]	2.232]	2.205]	2.320]	2.278]
Single HH	0.757***	0.803***	0.781***	0.812***	0.857**	0.881*	0.866**	0.882*
	[0.662;	[0.700;	[0.682;	[0.708;	[0.745;	[0.764;	[0.753;	[0.766;
	0.864]	0.921]	0.894]	0.932]	0.987]	1.015]	0.995]	1.016]
Care leaver	0.626***	0.650**	0.666**	0.686**	0.703*	0.715*	0.732	0.742
	[0.442;	[0.458;	[0.464;	[0.479;	[0.483;	[0.491;	[0.498;	[0.505;
	0.885]	0.923]	0.955]	0.982]	1.022]	1.039]	1.077]	1.089]
Potential FiF		0.494***		0.550***		0.669***		0.703***
		[0.429;		[0.474;		[0.578;		[0.604;
		0.569]		0.639]		0.775]		0.819]
Intercept	1.181***	1.916***	0.585	1.018	0.264***	0.376***	0.135***	0.200***
	[1.061;	[1.666;	[0.301;	[0.559;	[0.180;	[0.249;	[0.063;	[0.095;
	1.314]	2.202]	1.137]	1.854]	0.388]	0.566]	0.292]	0.422]
Further control variables								
Individual early educational attainment	No		No		Yes		Yes	
Gender	No		No		Yes		Yes	
School controls	No		Yes		No		Yes	

No. of observations: 5,047. The coefficients in each column are estimated in separate logit models to predict the probability of HE graduation controlling for missing flags. 95% confidence intervals based on robust standard errors clustered by sampling schools are in brackets. Coefficients are odds ratios. *** p<0.01, ** p<0.05, * p<0.1. Individual early educational attainment: capped linear KS4 (GCSE) score quintiles and total Key stage 2 score quintiles. School controls: the average KS2 and KS4 test scores of individuals' peers in the sampling secondary schools, secondary school type. Weighted using final Wave 8 weights. Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

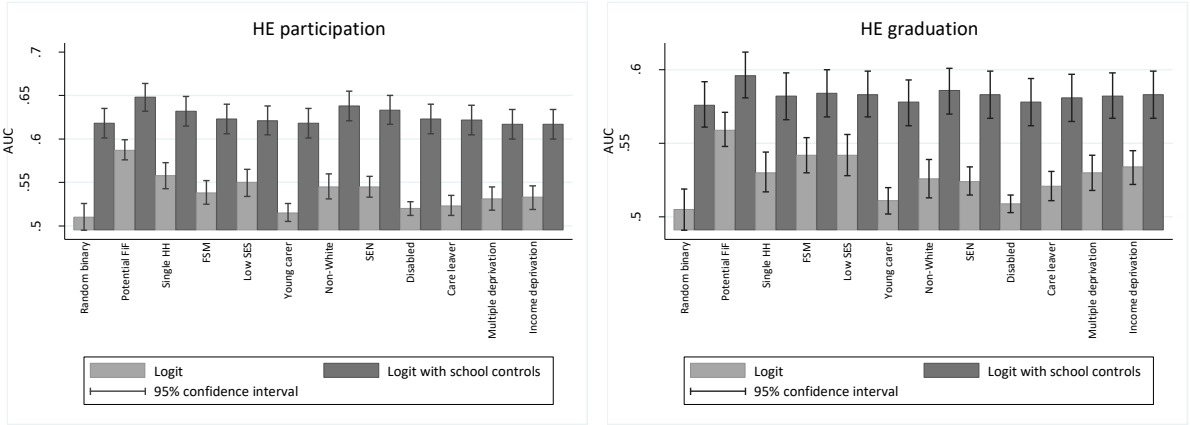
Figure 1. The distribution of the sample in terms of the number of disadvantages young people face



Notes: Complete cases only. N= 3,880. Unweighted. Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

Figure 2. The power of WP measures in terms of predicting the probability of HE participation and graduation in logit models with no with individual-level controls for early educational attainment: area under the ROC curve (AUC)

(a) **(b)**

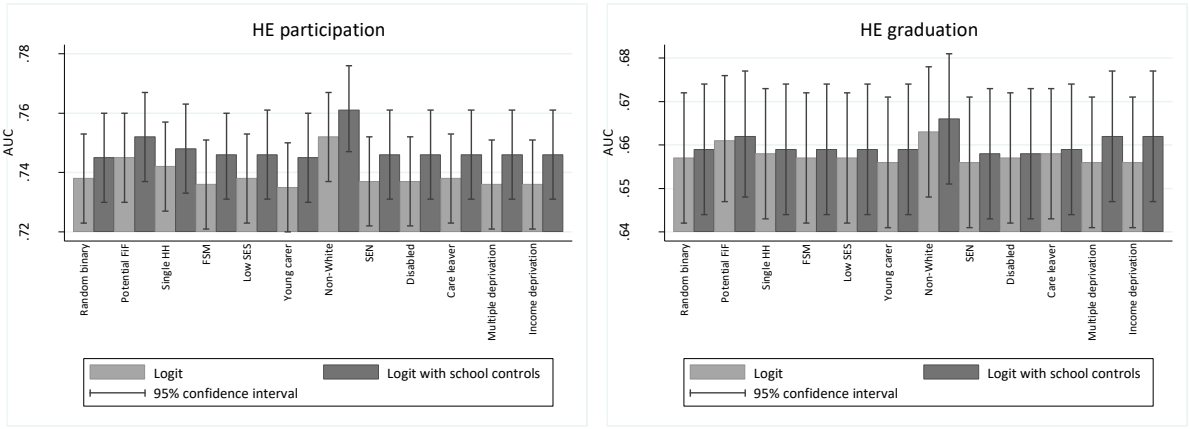


Notes: All AUC measures are estimated from separate equations controlling for missing flags. A random binary variable with mean=0.5 was also included to serve as a point of comparison. No. of observations: 5,047. The estimated logit coefficients from these logit models are reported in Table 1. School controls: the average KS2 and KS4 test scores of individuals' peers in the sampling secondary schools, secondary school type. Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

Figure 3. The power of WP measures in terms of predicting the probability of HE participation and graduation in simple logit models with individual-level controls for early educational attainment: area under the ROC curve (AUC)

(a)

(b)

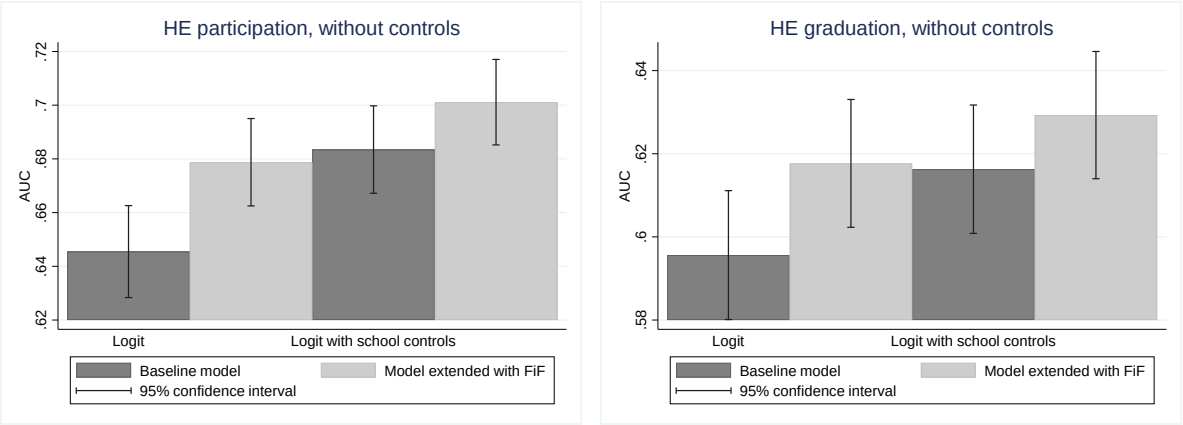


Notes: All AUC measures are estimated from separate equations controlling for missing flags. A random binary variable with mean=0.5 was also included to serve as a point of comparison. No. of observations: 5,047. The estimated logit coefficients from these logit models are reported in Table 2. Control variables: individual level KS2 and KS4 (GCSE) total score quintiles, gender. School controls: the average KS2 and KS4 (GCSE) test scores of individuals' peers in the sampling secondary schools, secondary school type. 'Low SES' refers to Low NS-SEC. Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. . DOI: 10.5255/UKDA-SN-7104-4

Figure 4. The additional predictive power of FiF when added to the parsimonious model in predicting the probability of HE participation and graduation in multiple logit models: area under the ROC curve (AUC)

(a)

(b)



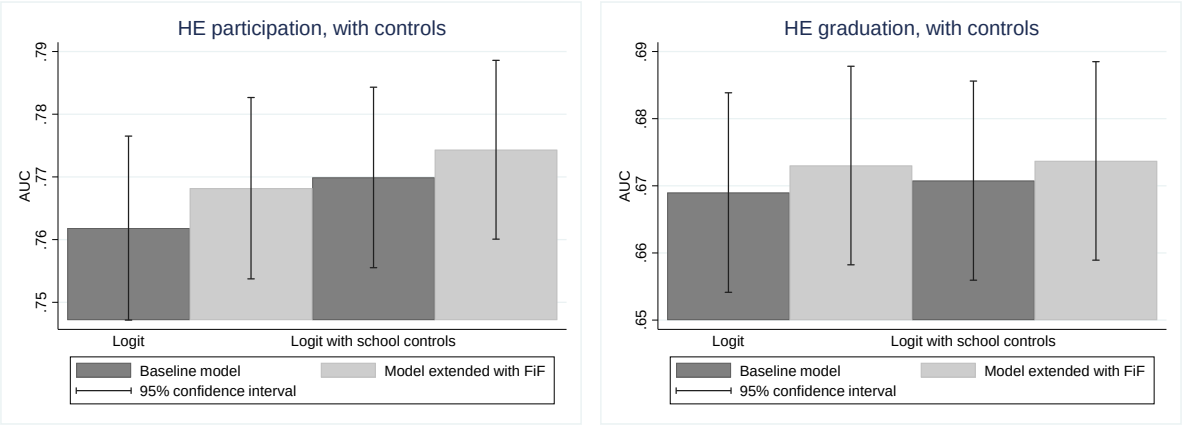
N=5,047. Explanatory variables in the baseline models: SEN, FSM, social class, income deprivation, non-white, disabled, single HH, care leaver, all with missing flags. School controls: the average KS2 and KS4 (GCSE) test scores of individuals' peers in the sampling secondary schools, secondary school type. The difference between the AUC of the baseline and the extended models has been tested using two-sided t-tests. In the logit models, both differences are significant (HE participation: p=0.0058; HE graduation: p=0.0478) while in the models including school controls they

are not (HE participation: p=0.1293; HE graduation: p=0.2425). Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

Figure 5. The additional predictive power of FiF when added to the parsimonious model in predicting the probability of HE participation and graduation in multiple logit models models with individual-level controls for early educational attainment: area under the ROC curve (AUC)

(a)

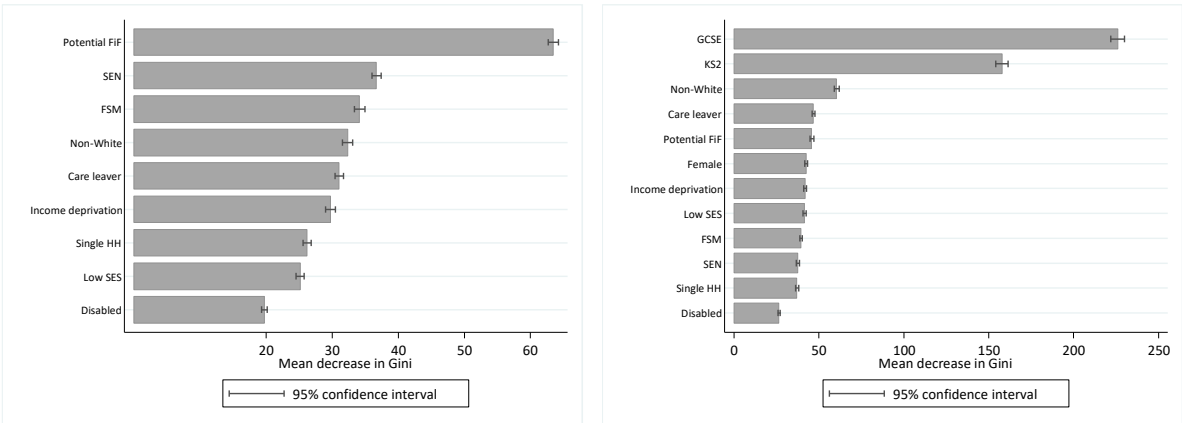
(b)



N=5,047. Explanatory variables in the baseline model: SEN, FSM, social class, non-white, single HH, care leaver, all with missing flags. Control variables: individual level KS2 and KS4 (GCSE) total score quintiles, gender. School controls: the average KS2 and KS4 (GCSE) test scores of individuals' peers in the sampling secondary schools, secondary school type. Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

Figure 6. The importance of predictors in predicting HE participation in random forest models: the estimated Mean Decrease in Gini coefficients of the predictors

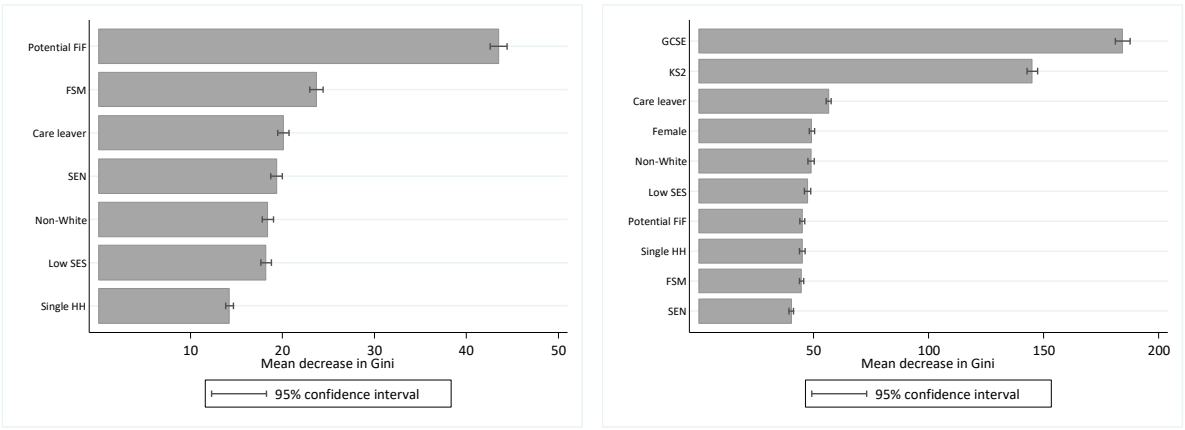
(a) **(b)**



N=5,047. Generated using the randomForest package in R. Missing values of the variables are controlled for by using missing flags. 'Low SES' refers to Low NS-SEC. The confidence intervals around the estimated parameters are constructed via bootstrapping (n=100). Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4.

Figure 7. The importance of predictors in predicting HE graduation in random forest models: the estimated Mean Decrease in Gini coefficients of the predictors

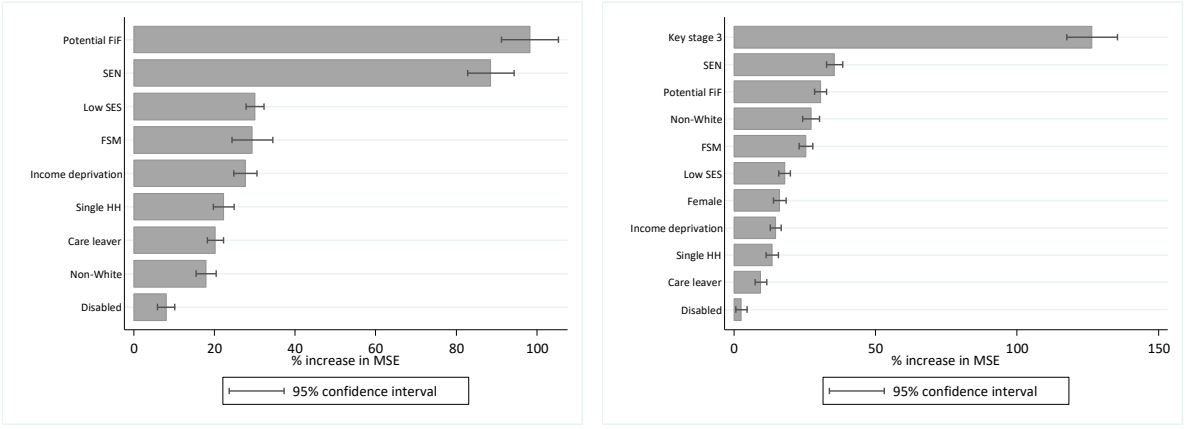
(a) **(b)**



N=5,047. Generated using the randomForest package in R. Missing values of the variables are controlled for by using missing flags. 'Low SES' refers to Low NS-SEC. The confidence intervals around the estimated parameters are constructed via bootstrapping (n=100). Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

Figure 8. The importance of predictors in predicting standardised capped linear GCSE scores in random forest models: percent increase in MSE

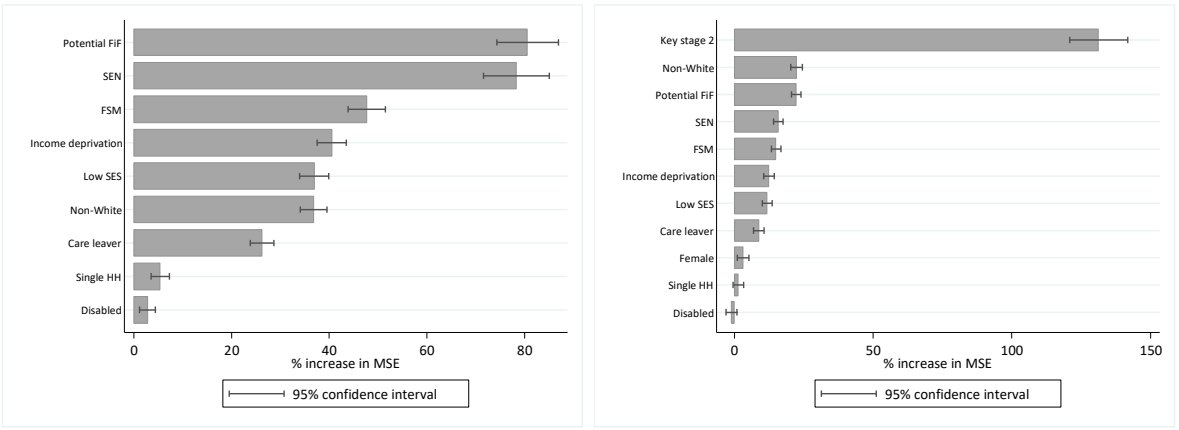
(a) **(b)**



N=4,360. Generated using the randomForest package in R. Missing values of explanatory variables are controlled for by using missing flags. 'Low SES' refers to Low NS-SEC. The confidence intervals around the estimated parameters are constructed via bootstrapping (n=100). Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

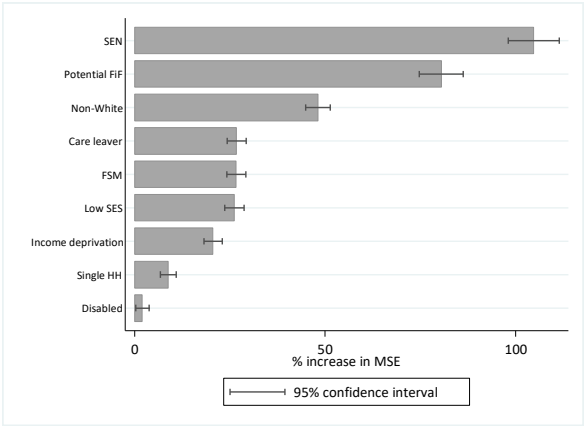
Figure 9. The importance of predictors in predicting standardized Key Stage 3 total point scores in random forest models: percent increase in MSE

(a) **(b)**



N=4,699. Generated using the randomForest package in R. Missing values of explanatory variables are controlled for by using missing flags. The confidence intervals around the estimated parameters are constructed via bootstrapping (n=100). Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

Figure 10. The importance of predictors in predicting standardised Key Stage 2 total point scores in random forest models: percent increase in MSE



N=4,613. Generated using the randomForest package in R. Missing values of explanatory variables are controlled for by using missing flags. The confidence intervals around the estimated parameters are constructed via bootstrapping (n=100). Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

Appendix

A. Next Steps compared to HESA data

In this paper we do not consider non-degree level HE qualifications, i.e. below-degree level HE diplomas, certificates, etc., neither in the case of the cohort members as mentioned above nor in the case of their parents. Thus, our definition of parental graduation (and first in family) differs from other data sources, in particular, the definition of the HESA Student Record data. The HESA Student Record data cover all UK domiciled students earning a degree and collects data on parental education as recorded on the UCAS application form when individuals apply to university (HESA, 2016). The following question is asked: "*Do any of your parents (as defined above) have any higher education qualifications, such as a degree, diploma or certificate of higher education*?". Due to the less strict definition of parental graduation in the HESA data, the share of first in family is considerably lower than in the Next Steps. Looking at a sample of first-degree graduates who are the same age as the Next Steps cohort and started their HE studies in 2008/2009 (without a gap year), the share of first in family graduates is 45%; in Next Steps it is 68.2% (Table 2). If we broadened our definition of graduated parents to those having either a university degree or any below-degree level HE diploma or certificate, the weighted share of FiF would be 49% among graduates, a similar share than in the HESA data (Table A1). The remaining four percentage points difference might be due to other dissimilarities in the two datasets: in Next Steps, individual might also have graduated after 2010/11 and might also have started their HE studies later than 2008/2009; the HESA data covers international students while Next Steps does not; parental education is reported by students in the HESA data while it is reported by parents in Next Steps; and lastly, information on parental education is missing in the HESA data in 21% of the cases while in Next Steps it is missing in only 0.7% of the cases.

Table A1: The share of first in family among graduates in Next Steps and HESA data

	Share of first in family (FiF) among graduates	Definition of parental graduation	Other features of the data collection
		Next Steps*	
Next Steps, unweighted	69.4	University degree (BA/BSc or higher)	Cohort: graduates born in 1989/90 in England. Full-time and part-time students. Parental education reported by parents. No. of observations: 2,671. Missing parental education on the top of this: 18 (0.7%).
Next Steps, weighted	68.2		
Next Steps, unweighted, HESA definition	50.2%	University degree or below-degree level HE certificates, diploma, etc.	
Next Steps, weighted, HESA definition	49.0%		
		HESA Student Records 2010/2011**	
HESA, weighting is not applicable	45.0%	University degree or below-degree level HE certificates, diploma, etc.	Cohort: UK domiciled graduates who started university in 2008/2009 and earned a first degree in 2010/2011 at ages 20/22. Full-time students only. Parental education is reported by the individuals. No. of observations: 107,926. Missing parental education on the top of this: 28,959 (21%).

Sources: *University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016; ** HESA Student Records 2010/2011. Sample restricted to UK domiciled full-time first-degree qualifiers 2010/11 with commencement date in the academic year 2008/2009.

B. Modelling strategy

Since schools are the main sampling units of Next Steps, and individuals were sampled within the chosen schools, we have to take account for the fact that individuals might have faced both school and individual (random) shocks, or in other words, the random shocks affecting individuals are expected be correlated within schools. Clarke et al. (2015) discuss the use of fixed effect (FE) vs. random effect (RE) modelling frameworks in similar setups, while Ilie, Sutherland, and Vignoles (2017) apply a multilevel (school and individual-level) random effect (intercept) modelling strategy to examine the validity of the FSM measure.

We choose not to apply a FE or RE framework in this case for two main reasons. First, our analytical sample is restricted to those with Level 3 education and includes 5,047 observations that either belong to 647 sampling schools chosen in the first wave of the study (4,944 individuals) or to the boost sample added to the survey in Wave 4 (103 individuals). The average number of observations per school is 7.6 and it varies between 1 and 23 per school.[16] Almost 60% of the sample belong to schools where the number of observations is at most 10. In 10% of schools, at least one of our outcome variables do not vary, and in 18%, our main variable of interest, first in family, perfectly predicts the outcome variable. Thus, the low number of observations by schools does not support the use of a FE or RE logit model. Second, students do not randomly choose their secondary schools and this choice is related to socioeconomic background (Anders 2012). If students from lower socioeconomic backgrounds were more likely to end up in particular schools at age 13/14, at the time of the initial sample selection, controlling for between-school differences in a FE or RE framework would already capture some of the effects of social background. More formally, random effect models would assume that school level errors are independent of the explanatory variables (Anders 2012) and this assumption would fail in our case. Similarly, in a FE framework, if school choice is already the consequence of socioeconomic characteristics, or is correlated with unobserved factors behind HE success, their inclusion as bad controls introduces a bias into the model (Angrist and Pischke 2008). Furthermore, in a FE or RE framework we can only estimate the within-school differences arising in the probability of HE participation and graduation due to social background characteristics, and we are not certain whether within-school differences or between- and within school differences together are more interesting for WP policy.

[16] Source: own calculation from University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016.

C. Lasso results

Table A2 presents the results of the Lasso procedure used to create the parsimonious model used in the second part of our empirical strategy. This table provides a comparison of how the Lasso treats certain variables as compared to the multivariate logistic regression. The variables with a zero in the Lasso column are those deemed unimportant in the system and therefore excluded from the parsimonious model.

Table A2. Lasso results (logit models, odds ratios) – HE participation and graduation

	HE participation		HE graduation	
	logit	Lasso	logit	Lasso
SEN	0.534***	0.613	0.698***	0.836
	[0.434; 0.656]		[0.572; 0.850]	
FSM	0.751**	0.835	0.783*	0.914
	[0.569; 0.992]		[0.605; 1.015]	
Low social class	0.672***	0.755	0.752***	0.859
	[0.573; 0.787]		[0.654; 0.864]	
Income deprivation	0.761	0.906	0.777*	1.000
	[0.543; 1.066]		[0.578; 1.045]	
Young carer	0.846	1.000	0.958	1.000
	[0.656; 1.092]		[0.760; 1.207]	
Non-White	2.203***	1.699	1.551***	1.237
	[1.794; 2.707]		[1.313; 1.832]	
Disabled	0.763*	0.926	0.939	1.000
	[0.554; 1.053]		[0.692; 1.275]	
Single HH	0.632***	0.701	0.809***	0.903
	[0.541; 0.738]		[0.704; 0.930]	
Care leaver	0.597***	0.770	0.647***	0.884
	[0.429; 0.831]		[0.469; 0.892]	
Multiple deprivation	0.918	1.000	1.094	1.000
	[0.667; 1.263]		[0.830; 1.442]	
Intercept	3.804***	1.000	1.303***	1.000
	[3.355; 4.312]		[1.173; 1.449]	

Notes: Complete case analysis, sample of those with Level 3 education. N=3,880. 95% confidence intervals based on robust standard errors clustered by sampling schools are in brackets. Coefficients are odds ratios. * p<0.05, ** p<0.01, *** p<0.001. Note that the Lasso procedure does not provide standard errors. Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

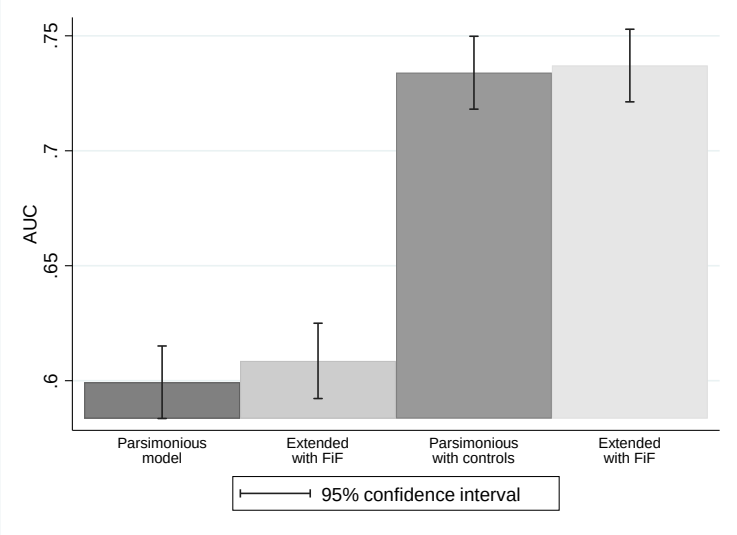
D. Robustness checks

D.1 Adding potential FiF on top of a parsimonious set of WP measures in random forest classification models

While the main reason for using the random forest classification algorithm is its built-in convenience to produce the ranking of predictors in terms of their importance even if they are not independent, it is straightforward to show that its predictions are comparable to those of the multiple logit models (subsection 4.2) and adding potential FiF on a set of WP measures leads to a similar increase in the predictive power of models using both approaches.

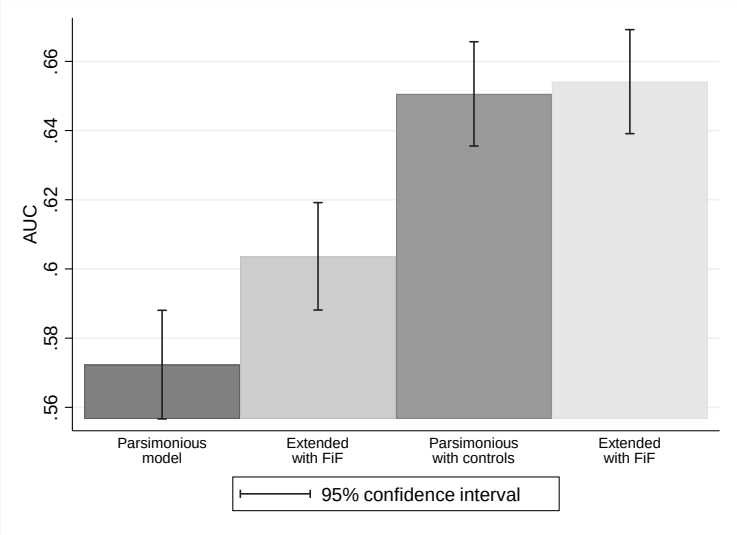
Following our earlier logic presented in subsection 4.2 for multiple logit models, we estimate four random forest classification models for both outcome variables: parsimonious model; parsimonious model extended with FiF; parsimonious model controlling for individual level early educational attainment; parsimonious model controlling for individual level early educational attainment, extended with FiF. Figures A1 and A2 compare the predictive power of the models using the AUC measure that is directly comparable across different specifications of logit and random forest models because it is estimated simply based on the share of correct predictions.

Figure A1. The additional predictive power of FiF when added to the parsimonious model in predicting the probability of HE participation in random forest models: area under the ROC curve (AUC)



N=5,047. Generated using the randomForest and the pROC packages in R. Explanatory variables in the baseline model: SEN, FSM, social class, income deprivation, non-white, disabled, single HH, care leaver, all with missing flags. Additional controls in the parsimonious models with controls: early educational attainment (KS3 total score quintiles and capped linear GCSE quintiles) and gender. Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

Figure A2. The additional predictive power of FiF when added to the parsimonious model in predicting the probability of HE graduation in random forest models: area under the ROC curve (AUC)



N=5,047. Generated using the randomForest and the pROC packages in R. Explanatory variables in the parsimonious model: SEN, FSM, social class, non-white, single HH, care leaver, all with missing flags. Additional controls in the parsimonious models with controls: early educational attainment (KS3 total score quintiles and capped linear GCSE quintiles) and gender. Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

We find similar results to those of the multiple logit models: without controlling for early educational attainment, adding FiF somewhat increases the predictive power of the models, considerably more in the case of predicting HE graduation (Figure A1) than in the case of predicting HE participation (Figure A2). Once early educational attainment is controlled for, however, any improvements become negligible. The predictive power of the logit and random forest models is also similar, it is around AUC=0.6 and 0.7 in both approaches, and below the rule-of-thumb threshold for good predictive power (AUC=0.8). The predictive power of the multiple logit models controlling for individual early attainment is the highest; however, the predictive power of logit models is based on in-sample predictions, while the predictive power of random forest classification models is based on out-of-sample predictions.

D.2 Predicting early educational attainment in linear models

We have shown in random forest regressions that potential FiF is an important predictor of early educational attainment. We emphasise that these results are not specific to the chosen method and traditional linear regressions would show a similar picture; although, we would not want to draw conclusions based on them due to the problems discussed earlier. Still, we show what happens if we estimate the same models that predict early educational attainment in subsection 4.4 in linear regressions estimated by OLS. We estimate the following models:

- we predict age 16 (GCSE) scores using WP measures (Model 1A);

- we predict age 16 (GCSE) scores using WP measures and controlling for age 13 (KS3) scores (Model 1B);
- we predict age 13 (KS3) scores using WP measures (Model 2A);
- we predict age 13 (KS3) scores using WP measures and controlling for age 11 (KS2) scores (Model 2B);
- we predict age 11 (KS2) scores using WP measures (Model 3).

Table A3. Disadvantage and early educational attainment (OLS, standardised outcomes)

	Outcome				
	Age 16 test scores (GCSE)		Age 14 test scores (KS3)		Age 11 test scores (KS2)
	Model 1A	Model 1B	Model2A	Model 2B	Model3
SEN	-0.673***	-0.296***	-0.656***	-0.185***	-0.810***
	[-0.777; -0.569]	[-0.374; -0.218]	[-0.748; -0.564]	[-0.258; -0.112]	[-0.906; -0.714]
FSM	-0.188**	-0.051	-0.225***	-0.114**	-0.242***
	[-0.333; -0.043]	[-0.167; 0.065]	[-0.348; -0.102]	[-0.212; -0.016]	[-0.393; -0.091]
Low social class	-0.095***	-0.050*	-0.142***	-0.093***	-0.075**
	[-0.156; -0.034]	[-0.101; 0.001]	[-0.209; -0.075]	[-0.152; -0.034]	[-0.132; -0.018]
Income deprivation	-0.091*	0.001	-0.143***	-0.161***	-0.013
	[-0.189; 0.007]	[-0.073; 0.075]	[-0.239; -0.047]	[-0.237; -0.085]	[-0.119; 0.093]
Non-White	-0.006	0.115***	-0.118***	0.063*	-0.265***
	[-0.082; 0.070]	[0.054; 0.176]	[-0.198; -0.038]	[0.000; 0.126]	[-0.353; -0.177]
Disabled	-0.339***	-0.225***	-0.292***	-0.191***	-0.236***
	[-0.519; -0.159]	[-0.384; -0.066]	[-0.447; -0.137]	[-0.322; -0.060]	[-0.399; -0.073]
Single HH	-0.160***	-0.108***	-0.079***	-0.036*	-0.065**
	[-0.215; -0.105]	[-0.151; -0.065]	[-0.130; -0.028]	[-0.077; 0.005]	[-0.120; -0.010]
Care leaver	-0.073	-0.026	-0.002	0.088*	-0.129*
	[-0.198; 0.052]	[-0.120; 0.068]	[-0.122; 0.118]	[-0.014; 0.190]	[-0.260; 0.002]
Potential FiF					
	-0.349***	-0.140***	-0.314***	-0.115***	-0.328***
	[-0.402; -0.296]	[-0.181; -0.099]	[-0.375; -0.253]	[-0.164; -0.066]	[-0.385; -0.271]
N	4360	4360	4699	4699	4613
	Control variables				
Earlier educational attainment	No	KS3	No	KS2	No
Gender	Yes	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes	Yes

Coefficient of all columns are estimated in separate liner regression models to predict the standardised values of test scores, controlling for missing flags. 95% confidence intervals based on robust standard errors clustered by sampling schools are in brackets. Coefficients are interpreted in the standard deviation of the dependent variable. *** p<0.01, ** p<0.05, * p<0.1. School controls: the average KS2 and KS4 test scores of individuals' peers in the sampling secondary schools, secondary school type. Weighted using final Wave 8 weights. Source: University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2018). Next Steps: Sweeps 1-8, 2004-2016: Secure Access. DOI: 10.5255/UKDA-SN-7104-4

Our results are presented in Table A3. Potential FiF is a significant negative predictor of early educational attainment at all ages even after controlling for all other WP measures. Interestingly, the statistical relationship between potential FiF and standardised test scores measured at age 16,

14 and 11 are very similar (-0.349, -0.314 and -0.328, respectively). Once earlier educational attainment is controlled for (Model 1B and 2B), the magnitude of its coefficient decreases but stays highly significant between -0.115 and -0.140. While this exercise is not useful in terms of ranking the importance of the predictors of early educational attainment (note that comparing the magnitude of the estimated coefficients is not informative), it supports our earlier results that potential FiF is a systematic barrier in early schooling achievements even after controlling for other measures of disadvantage and earlier educational achievement.