Revisiting the returns to higher education: heterogeneity by cognitive and non-cognitive abilities

Oliver Cassagneau-Francis*

19th May 2022
(click here for the latest version)

Abstract

Recent work has highlighted the significant variation in returns to higher education across individuals. We develop a novel methodology—exploiting recent advances in the identification of mixture models—which groups individuals according to their prior ability and estimates the wage returns to a university degree by group. We prove the non-parametric identification of our model. Applying our method to data from a UK cohort study, our findings reflect recent evidence that skills and ability are multidimensional. Our flexible model allows the returns to university to vary across the (multi-dimensional) ability distribution, a flexibility missing from commonly used additive models, but which we show is empirically important. The returns to higher education are 3–4 times larger than the returns to prior cognitive and non-cognitive abilities. Returns are generally increasing in ability for both men and women, but vary non-monotonically across the ability distribution.

Keywords: Mixture models; Distributions; Treatment effects; Higher education; Wages; Human capital; Cognitive and non-cognitive abilities.
JEL codes: E24; I23; I26; J24

*Sciences Po, Paris; email: oliver.cassagneaufrancis@sciencespo.fr. I am grateful to my advisors Ghazala Azmat and Jean-Marc Robin for their invaluable guidance and support throughout this project. I thank Stéphane Bonhomme, Moshe Buchinsky, Eric French, Kerstin Holzheu, Hugo Lhuillier, Stefan Pollinger, Weilong Zhang, and numerous seminar participants at Sciences Po and the University of Cambridge for helpful comments and discussions. I thank the Centre for Longitudinal Studies in the IOE at University College London for administering the survey and, along with the UK Data Service, providing access to the data. The usual disclaimer applies.

1 Introduction

Economists have long concerned themselves with estimating the wage returns to education, beginning (at least) with Mincer in 1958. For almost as long, critiques of this work have argued that a failure to control for ability has led to a significant "ability bias" in estimates of the returns to education. These claims, whether well-founded or not (Griliches, 1977), led to attempts to circumvent the ability bias problem using instrumental variables. However, these estimates were *higher* than the ability-biased estimates, which were themselves supposed to be biased upwards. This prompted a new approach, recognising that the returns to education are not homogeneous, and so using different methods and instruments would lead to different estimates (Blundell, Dearden, and Sianesi, 2005).

By allowing heterogeneous returns to higher education, we join this growing literature which explicitly studies the variation in returns to a university degree across across individuals and groups of individuals. The variation in returns across different groups can be considerable.[1] However, there is little evidence on how the returns to education vary with a fundamental characteristic, ability.[2] Ability is important not only as a potential source of bias; lessons from the literature on human capital formation[3] suggest a person's ability is also likely to directly impact the returns they achieve from a university degree. This paper develops and estimates a framework explicitly designed to investigate how the returns to university vary flexibly with what we call "prior ability": a young person's cognitive and non-cognitive ability on entry to university.

Our focus on cognitive and non-cognitive ability recognises the growing body of evidence that skills are multidimensional, and that collapsing these dimensions into a single (usually cognitive) measure misses important sources of variation across people.[4] Our analysis

[1] Britton, Dearden, and Waltmann (2021a) investigate how the returns to university vary across socio-economic and ethnic groups in the UK. They find positive returns to university for all groups, though substantial heterogeneity: returns are higher for women than for men, and across ethnic groups they vary from 7% for White British men, to 40% for Pakistani women. Britton, van der Erve, Belfield, Dearden, Vignoles, Dickson, Zhu, Walker, Sibieta, and Buscha (2021b) study the returns to different subjects and institutions, again finding substantial heterogeneity in the returns to different subject and institutions after controlling for prior cognitive ability. They find weak evidence that returns are positively correlated with the selectivity of the subject or institution.

[2] We use the terms "human capital", "skills", and "ability" interchangeably throughout this paper.

[3] Recent work by Cunha, Heckman and coauthors (2006; 2007a; 2008; 2010) has shown that skills obtained early in childhood are vital for fostering skills later in childhood—a feature they call the *self-productivity* of skills. A related concept is the *complementarity* of skill formation: "skills produced at one stage raise the productivity of investment [in further skills] at subsequent stages" (Cunha et al., 2006, p. 703). These features of skill production during childhood suggest that ability on entry to university will affect the impact of a university degree on an individual's ability, and hence on their later outcomes.

[4] Focusing on educational outcomes, Jacob (2002) finds that non-cognitive skills are key in explaining the gender gap in college attainment in the US, and Delaney, Harmon, and Ryan (2013) demonstrate the link between non-cognitive skills and study behaviours known to be important for success in undergraduate degrees. Turning to success later in life, Heckman, Stixrud, and Urzua (2006) offer evidence that non-cognitive skills are important for a range of social and economic outcomes. Bowles, Gintis, and Osborne (2001) survey the literature on the determinants of earnings, with a particular focus on non-

incorporates these insights by allowing both cognitive and non-cognitive skills to determine wages, and hence the wage returns to a university degree. We compare results obtained using only cognitive measures with those using both cognitive and non-cognitive measures, thereby providing further evidence of the important role for non-cognitive ability.

A key contribution of our paper is methodological: we adapt the framework of Cassagneau-Francis, Gary-Bobo, Pernaudet, and Robin (2021), where we proved the non-parametric identification of, and developed an estimation strategy for, a model of formal training and wages.[5] This work continues a long tradition in economics of using discrete mixtures to model heterogeneity, going back to Heckman and Singer (1984). In recent years, major progress has been made in the identification of this type of model, and in the development of nonparametric estimators (see for example Bonhomme, Jochmans, and Robin, 2016a and the references therein). Here and in Cassagneau-Francis et al. (2021) we make novel attempts at using discrete mixtures in the context of evaluation models.

Our statistical model, motivated by the human capital formation literature mentioned above, builds upon the work of Heckman and coauthors (2003; 2006; 2007a; 2010). In these papers, the analysis typically requires strong functional form assumptions to identify the model, often assuming an underlying factor model structure (Carneiro et al., 2003).[6] By adapting the framework in Cassagneau-Francis et al. (2021), we are able to achieve identification of a yet more flexible model, estimate a non-linear version of our model, and demonstrate the importance of these non-linearities empirically.[7]

In order to identify our non-linear model, we assume that the distribution of prior ability (i.e. of latent types in our framework) is discrete. Our method is frugal in its requirements of the data available to the econometrician, and having discrete types means our heterogeneity analysis is easily interpreted. The costs of this flexibility, frugality, and interpretability are low: (i) we require, in addition to a measurement of each component of ability and an outcome,[8] a single crude (i.e. discrete) measurement of ability, *or* a discrete instrument for university attendance (though exogeneity is only required conditional on type), and (ii) we assume the distribution of prior ability has finite support.

cognitive traits. More recently, Todd and Zhang (2020) include personality traits in a dynamic discrete choice model of schooling and occupational choice. The authors find important links between personality and schooling, and between personality and occupational choice.

[5]Similar techniques have been used to identify a range of models including: firm and worker sorting (Bonhomme, Lamadon, and Manresa, 2019); and the contributions of workers across different teams (Bonhomme, 2021). Gary-Bobo, Goussé, and Robin (2016) identify and estimate a parametric model, also inspired by the human capital formation literature, to study the effects of grade retention on French middle school students.

[6]Cunha et al. (2010) show how to relax some of the stronger assumptions, allowing the measurement and outcome equations to be non-linear in their inputs.

[7]Cunha et al. (2010) only estimate the additive version of their model.

[8]Cunha and Heckman and coauthors require at least two measurements per component, plus an outcome.

Following Cassagneau-Francis et al. (2021), we show the non-parametric identification of the returns to university conditional on an individual's prior ability, which in our model is summarised by their latent type. We can aggregate these type-conditional "treatment effects" (TE) to obtain more standard effects: average treatment effect (ATE), average treatment on the treated (ATT), and Imbens and Angrist (1994)'s local average treatment effect (LATE). We can also aggregate the type conditional TEs to obtain analogues of the usual OLS and IV estimates, which we show to be biased estimators of ATT (for OLS) and LATE (for IV). Having shown non-parametric identification, we adapt our estimation strategy from Cassagneau-Francis et al. (2021), specifying a parametric specification of our model. This approach allows us to use a sequential version of Dempster, Laird, and Rubin (1977)'s expectation-maximisation (EM) algorithm. We use a bootstrap procedure to calculate standard errors.

In our application, we use our framework to estimate the returns to a university degree in the UK as a function of cognitive *and* non-cognitive prior ability. Our data come from the British Cohort Study (BCS 1970), which follows all individuals born in the UK in a single week in 1970, and contains detailed information on the cohort members at age 16 (before attending university) and again at age 26 (after university). In particular, the young people took cognitive tests and answered a series of questions to capture their non-cognitive abilities at age 16. Crucially we also observe their wages at age 26, along with any qualifications they have achieved up to that point and hence whether they graduated from university.

We estimate our model separately by gender,[9] a data-driven decision resulting in better performance from our estimation algorithm. Although the graduation rates for men and women are quite similar (33% for men, and 27% for women), there was a large gender wage gap during this period (reflected in our data), both for graduates and non-graduates. Our algorithm struggled to deal with this difference when we pooled genders.[10] A large and important literature attempts to uncover the institutional and societal factors driving this gap, a task which is beyond the scope of this paper.

We find that the returns to a university degree for our UK cohort are generally positive and large for both men (10–20%) and women (15–28%), but vary significantly with prior ability—i.e. across individuals of different types in our framework. This variation is also highly non-linear, with the size of the effect varying non-monotonically across the ability distribution. However, these patterns are quite different across genders. For men, the returns are U-shaped with respect to prior ability, with middle-ability types receiving

[9] Blundell, Dearden, Goodman, and Reed (2000) also estimate their model separately by gender, and study a similar UK cohort born 12 years earlier than our cohort, though consider wages at 33, so 5 years earlier than we observe wages.

[10] Estimating our model separately across genders is the most straightforward way to "control for" gender in our framework.

the lowest returns. The opposite is true for women, for whom we observe hump-shaped returns with the highest for middle-ability types.

This non-linearity would not be apparent under the current leading estimation approaches which assume an additive model for wages. In a linear (additive) model, the wage returns to university are proportional to a young person's ability level. Therefore, if the returns to university are increasing in ability *on average*, we would estimate a higher return for a high-ability young person than a low-ability young person—even if this relationship only holds for part of the ability distribution.

Our analysis also reveals that the returns to a university degree in the UK are more important than the returns to ability in the following sense: a low-ability young person can earn higher wages by completing university, and becoming a low-prior-ability graduate, than they could by improving their ability to become a high-ability non-graduate. The large impact of university on wages across the ability distribution drives another of our main results: the contribution of the graduate wage premium to wage inequality is 3 (men) and 4 (women) times larger than the contribution of ability attained *prior to university*. Our results complement those of Cawley, Heckman, and Vytlacil (2001) who find that, having controlled for educational attainment, cognitive ability explains little of the variation in wages across individuals even within occupations. We find that *both* cognitive and non-cognitive skills explain only a small part of wage inequality.

Despite the relatively small *direct* contribution of ability to wage inequality, a young *man's* levels of cognitive and non-cognitive skills on entry to university do influence the returns they can expect to achieve. There is a significant comparative advantage for non-cognitive skills among male non-graduates, resulting in low returns to university for high non-cognitive-ability, middle-cognitive-ability men. The equivalent is not true for women. To what extent this is due to the different occupations favoured by male and female non-graduates remains a question for future research.

The remainder of the paper proceeds as follows. In section 2 we present the setup of our model, with a discussion of identification in 2.1. Section 3 describes how we estimate the model. We then turn to our application using UK-cohort data: estimating the wage returns to a university degree as a function of cognitive and non-cognitive prior ability. Section 4 discusses the relevant context of higher education in the UK, and presents our dataset and some initial descriptive results. Section 5 presents the results of estimating our model on this data, first with only cognitive ability, and then with both cognitive and non-cognitive components. Section 6 concludes.

1.1 Related literature

Returns to education. There is a long tradition in economics of attempts to estimate the returns to schooling, a tradition which perhaps began with the seminal work of Mincer (1958, 1974). Critiques of this early work suggested it was plagued by issues of ability bias, which although arguably small (Griliches, 1977), led to a search for sources of exogeneous variation and the use of IV methods to avoid this criticism. Card (1999) provides an excellent summary. However, despite these methods being used to avoid the positive ability bias, IV estimates of the returns to schooling are typically larger than OLS estimates. These apparently contradictory findings were due to either an even larger ability bias, for family background IVs, or particularly high *marginal* returns for those impacted by institutional IVs — Imbens and Angrist (1994)'s local average treatment effect, or LATE.

These high marginal returns estimated by IV methods highlighted another avenue to explore: the returns to education are unlikely to be constant, varying with both observed and unobserved characteristics. Initial work used sibling Altonji and Dunn (1996) and twin (Ashenfelter and Rouse, 1998) studies to analyse the effects of family background on the returns to education, finding little variation. Barrow and Rouse (2005), also focusing on siblings, find little effect of race and ethnicity on the returns to education.

Much of this work focused on the return to an additional year of schooling. Other authors have focused on the returns to educational milestones, with the returns to a university degree being most relevant for this paper (see Kane and Rouse (1995) for evidence from the US and Blundell et al. (2000) from the UK). Allowing returns to be heterogenous, Carneiro, Heckman, and Vytlacil (2011) estimate returns to college that vary with the unobserved cost of attaining a degree. A recent paper by Britton et al. (2021b) studies how the returns vary across different degrees and institutions, as well as across different socio-economic and ethnic groups. We join this growing literature on the heterogeneous returns to a university degree, estimating wage returns which vary with both cognitive and non-cognitive ability.

Another relevant strand of the literature compares the returns to ability with those to education. Taber (2001) argues that the growth in the wage premium in the US in the 1980s is largely driven by an increase in the demand for high-skill (i.e. college-education) workers. Cawley et al. (2001) find that cognitive ability explains only a small part of wages once schooling is controlled for, and highlight that non-cognitive ability is also important for labour marker outcomes.

Human capital formation and non-cognitive ability. The model in our paper is inspired by the literature on human capital formation. This literature, which mainly fo-

cuses on the production of human during childhood, contains a number of lessons relevant for our analysis. Early work on human capital distinguished it from "ability", as being something that something that could be invested in, unlike innate ability which was invariant (Becker, 1964). The focus was entirely on cognitive ability (and human capital). More recent work has argued that there is no true distinction between human capital and ability — whether called skills, ability, or human capital, these traits are a product of an individual's genes, environment, and can be acquired and improved; and that both cognitive and non-cognitive abilities are important, both for success during formative years by fostering further improvements in these abilities, and also for later outcomes. These findings are summarised by Cunha et al. (2006, henceforth CHLM) in an excellent review.

We borrow a number of insights from this literature. CHLM emphasise two key related features of skill formation which we also incorporate into our analysis: (i) skills produced at one stage of development are important for fostering skills at later stages, CHLM's "self-productivity" of skills; (ii) later investment in skills is necessary to fully realise the benefits of earlier investments — in CHLM's terminology the "complementarity" of skills. Our contribution is to embed these insights from childhood development into a model of investment in skills at a later stage of the life-cycle: higher education. We design a framework to study the returns to higher education, explicitly allowing returns to vary with prior cognitive and non-cognitive abilities. As far as we are aware, our paper is the first to estimate a model of this type allowing for non-linear measurements and outcomes.

Model, identification and estimation. Our empirical framework is close in spirit to the recent papers of Cunha and Heckman (2007a, 2008) and Cunha et al. (2010) who aim is to estimate the technology of skill formation. Similar to the aforementioned papers, we assume latent factors which link measurements and outcomes, and like Cunha et al. (2010) we are able demonstrate the non-parametric identification of our model.

We depart from this work in assuming a discrete distribution for these latent types. Similar assumptions have recently been used to model unobserved heterogeneity by a number of authors. Bonhomme and Manresa (2015) investigate the properties of using latent types (or groups) to capture (unobserved) heterogeneity, in a setting the authors call "group fixed effects". This type of setup is explored further in Bonhomme, Lamadon, and Manresa (2022). These methods have been applied to study matched employer-employee data (Bonhomme et al., 2019) and the contributions of individuals in team settings (Bonhomme, 2021).

Closest to our work are the papers by Gary-Bobo et al. (2016) and Cassagneau-Francis et al. (2021). Gary-Bobo et al. (2016) estimate the effects of grade retention on French middle school students, while Cassagneau-Francis et al. (2021) estimate the wage returns to formal training in France, both relying on discrete types to capture unobserved het-

erogeneity. Both these papers employ on a differences-in-differences-like setup, observing the same outcome before and after treatment — our paper adapts this method to allow different measurements/outcomes before and after treatment.

We follow the identification proof in Cassagneau-Francis et al. (2021), which relies on recent advances in the identification of finite mixtures, using matrix algebra to prove identification. We refer the interested reader to a series of papers by Bonhomme et al. (2016a); Bonhomme, Jochmans, and Robin (2016b, 2017) and citations therein, for further details. We use Dempster et al. (1977)'s EM algorithm to estimate our model, following both Gary-Bobo et al. (2016) and Cassagneau-Francis et al. (2021). This method has been used widely in economics, providing a relatively straightforward method to estimate models with "missing" data (our latent types). However, it can still be computationally intensive, involving maximising complex likelihood functions. Arcidiacono and Jones (2003) show how an alternative formulation of the problem allows *sequential estimation*, allowing parameters to be updated separately.

2 Empirical framework

There are N young people indexed by i. We denote their (log)-wage by w_i, observed at age 26 when they are either university graduates, denoted $d_i = 0$, or non-graduates ($d_i = 0$). Their wage depends on their ability before attending university, which we will call "prior ability" and denote by θ, and on whether they graduate from university. Ability is multi-dimensional. We focus on the two-dimensional case, in which individuals might differ in their cognitive (θ^C) and non-cognitive (θ^N) abilities. Then $\theta = (\theta^C, \theta^{NC})$. The different components of ability may or may not be correlated. Our aim is to estimate the causal impact of graduating from university on wages, as a function of prior ability.

However, we do not observe ability (θ) directly. Ability is the classic confounder in attempts to estimate the returns to university, being both a determinant of wages *and* of the decision to attend university (Becker, 1964; Card, 1999). It is also more fundamental to our analysis, given we want to study how the returns to university vary with prior ability. We follow the example of both the recent literature on human capital formation and on the returns to schooling (see for example Cunha and Heckman, 2007a; Carneiro et al., 2011), relying upon noisy measurements of a young person's prior ability. We have (at least) one measurement specific to each ability, i.e. a purely cognitive measurement and a purely non-cognitive measurement. Using the information on θ contained in these measurements and in wages, we are able to identify and estimate the distribution of θ, and hence study how the returns to university vary across this distribution.

We depart from this recent literature in how we model the dependence of measurements and wages on ability. Typically, authors assume mean measurements and wages are linear

functions of ability, with higher moments of the distribution independent of ability (see for example Carneiro et al., 2003). A linear version of our model is presented and briefly discussed in appendix A. We relax these assumptions, and imposing no functional form on how the means of these distributions depend on ability, and we can allow the variance to be a function of ability (though we do not in our application). To achieve this, we assume the distribution of prior ability has finite support. Under this assumption, we can classify individuals into a finite number of groups based on their prior ability, groups across which the distributions of measurements and wages vary systematically. We denote these groups by $k \in \{1,\ldots,K\}$. Therefore, ability takes only a finite number of values, which we can index by the group identifier, k, so that $\theta_k = (\theta_k^C, \theta_k^{NC})$.

In addition to the continuous wages and measurements there is a discrete variable z, which is either an additional (crude) measurement of ability, or an "instrument" for university graduation. We do not include any other control variables when discussing identification, or when estimating our model. However, adapting our proof and estimation strategy to include controls would be straightforward, though it would require additional restrictions on our model. We say "instrument" as z need only be independent of measurements and wages *conditional on prior ability*, i.e. on a young person's type. This idea is formalised in our first assumption.

Assumption 1 (Measurements and wages)**.** *Measurements, wages and z are independent conditional on type and education.*[11]

We denote the distribution of wages conditional on type and education by $f_w(w_i|k,d)$. Similarly, the conditional distribution of the measurements is $f_\ell(M_i^\ell|k,d)$, $\forall \ell \in \{C, NC\}$. The probability mass of young people of type $k \in \{1,\ldots,K\}$, with value of the instrument $z \in \{1,\ldots,Z\}$, and with education level $d \in \{0,1\}$, is denoted by $\pi(k,z,d)$.[12] We want to identify and estimate these objects, along with the distribution of prior ability $(\theta_k, \forall k = 1,...,K)$. Before discussing our identification strategy, we present briefly the economic foundations of our statistical model.

[11] Measurements need not be independent of each other even conditional on type.

[12] Our model is closely related to the extended Roy model (Heckman and Vytlacil, 2005; Carneiro, Heckman, and Vytlacil, 2010; Carneiro et al., 2011):

$$w = w(k,0) + [w(k,1) - w(k,0)]\, D$$
$$D = 1 \text{ if } \mathbb{E}[w(1) - w(0)|k] \geq c(k,z),$$

where k is an individual's type (capturing their cognitive and noncognitive ability). z is the instrument, i.e. an environmental variable affecting treatment decision, through the non-pecuniary cost of attending university, but independent of wages and measurements conditional on type. $w(k,0), w(k,1)$ are treatment-specific outcome variables (random given k and independent of z). $c(k,z)$ is cost of attending university (random given k,z).

Economic motivation for the model. Our statistical model is motivated by the literature on human capital formation. Consider a simple model of human capital formation (Todd and Wolpin, 2003; Cunha et al., 2006). There are two periods: before university ($t=0$, age sixteen in our application); and after university ($t=1$, age twenty-six in our application). Human capital (ability) in period $t+1$ is a function of human capital in the previous period, θ_t and any investments in human capital made between t and $t+1$, I_t.

$$\theta_{t+1} = f_\theta(\theta_t, I_t) \tag{1}$$

We can simplify the notation in equation (1) to include just this single period, between $t=0$ and $t=1$, and we additionally assume that investment in human capital during this period is binary: young people either attend and graduate from university or they do not. We can also use k and θ_0 interchangeably, so

$$\theta_1 = g_\theta(\theta_0, d) = g(k, d)$$

Then, if wages in $t=1$ are a function of human capital in $t=1$, we obtain our model for wages

$$w_i \sim \tilde{f}_w(w|\theta_1) = \tilde{f}_w(w|g(k,d)) = f_w(w|k,d).$$

2.1 Non-parametric identification

One of the key contributions of this paper is a novel identification and estimation strategy that does not rely on wages and measurements being linear in their components. Our strategy requires fewer measurements of prior ability for identification than the current leading factor-model approach, and we are able to identify and estimate a fully non-linear model.[13] This frugality and flexibility come at a low cost, as compared with the linear factor-model approach we additionally require: 1) a crude (can be binary or discrete) measurement of prior ability,[14] *or* a crude instrument for university attendance, and which need only be exogenous of wages and measurements conditional on prior ability, which we denote z; 2) that the distribution of prior ability has finite support.

We assume during our discussion of identification that the econometrician knows the *true* number of points of support, K. However, we also discuss how this can be estimated when we operationalise the method. Recall that our aim is to identify the discrete distribution of prior ability, θ_k and $\pi(k)$,[15] the distributions of measurements and wages conditional

[13] Cunha et al. (2010) prove identification of a non-linear version of the factor model, but rely on additively separable measurement and outcome equations when estimating their model. Their method requires the same number of observations (measurements) as the linear model.

[14] This measurement is not really *additional* to the factor model; we could use one of the extra measurements required in that approach, discretising the variable if it is continuous.

[15] Our method identifies $\pi(k,z,d)$ which we can then sum over z and d to obtain $\pi(k)$, the proportion

on prior ability and education, $f_\ell(M^\ell|k,d)$ and $f_w(w|k,d)$.

Individual measurements and wages are not directly informative of prior ability due to noise in these measures / outcomes. However, under our maintained assumptions, mean wages and measurements across individuals of the same type—who share the same prior ability, θ—*are* informative. Therefore we can use these means to identify the support of θ, i.e. θ_k, and to set it in an interpretable metric allowing comparisons *across* types.

Likelihood of an individual's observations. Under the model detailed at the start of section 2, the likelihood associated with an individual i's observations writes

$$\ell(\boldsymbol{M}_i, w_i, d_i, z_i) = \sum_k \pi(k, z_i, d_i) f_M(\boldsymbol{M}_i|k) f_w(w_i|k,d) \tag{2}$$

with $\boldsymbol{M} = (M^1, \ldots, M^L)$, $f_M(\boldsymbol{M}|k) = \prod_\ell f_\ell(M^\ell|k)$.

We follow Cassagneau-Francis et al. (2021) and apply results from recent work on mixture models to show the elements on the right-hand side of equation (10) are identified under certain conditions (see Bonhomme et al. (2017) for details). A formal statement of the necessary assumptions, our identification theorem and a detailed proof is in appendix B. Here, we only summarise the key assumptions and ideas of the proof.

We have already introduced one of the main assumptions (assumption 1); that measurements, wages and the instrument (if used) are independent conditional on type. This allows for dependence of higher moments of the measurement and wage distributions on latent types (i.e. on ability), not just the means of these distributions.[16] The key here is that all the dependence across wages and measurements is summarised by a person's type.

Next, we require the wage and measurement distributions (excluding z) to be continuous, or at least sufficiently granular that the type-conditional wage and measurement distributions are linearly independent. We cannot identify any latent types whose wage (measurement) distribution is a linear combination of the wage (measurement) distributions of the other types.

There are then two conditions on the probability mass function, $\pi(k,z,d)$. The role of z in the identification is to form similar systems, one for each value of z. These systems are similar in that they contain the same measurement and wage distributions, but we rely on z to ensure they are sufficiently different to allow identification of all the components.[17]

of individuals of type k.

[16] This might be important as one can imagine some young people being particularly good (or bad) at tests, a "skill" that would affect both their cognitive and non-cognitive scores, but one that might not neccessarily be valued by employers.

[17] We refer the reader to the proof in appendix B for formal notions of similar and different in this

For this, z either needs to be correlated with d, but not with wages (conditional on k) i.e. z is an "instrument"; or z needs to be correlated with k, i.e. z is an additional measurement; or both.

Finally, we need to label types consistently across values of d. Our method identifies $f(M|k,d)$, $f_w(w|k,d)$, and $\pi(k,z,d)$ separately for each value of d. But, there are no young people with all values of d that would allow us to label types consistently. Therefore we need another assumption. We assume that the measurement distributions are independent of education, conditional on type. Then, we can equate these distributions across values of d to label k consistently.

Having identified these distributions, we are able to calculate heterogeneous returns to university, conditional on prior ability. We are in a *potential outcomes* framework, with each individual having two potential outcomes, w_1 and w_0, of which we only observe one. The ATE under this framework is $\mathbb{E}[w_1 - w_0]$, although only $\mathbb{E}[w_1 \,|\, d=1]$ and $\mathbb{E}[w_0 \,|\, d=0]$ are observed. If we believe that wages are correlated with education, then

$$\mathbb{E}[w_1 - w_0] \neq \mathbb{E}[w_1 \,|\, d=1] - \mathbb{E}[w_0 \,|\, d=0]. \tag{3}$$

This is the classic problem of *selection into treatment*, where treatment in our case is a university degree.

Fortunately, our maintained assumptions permit a solution. Potential outcomes are independent of education and z conditional on prior ability, i.e. $w_1, w_0 \perp\!\!\!\perp z, d \,|\, k$. We can then define the following type-conditional "treatment effect".

Average treatment effect by type, $ATE(k)$. This is the expected wage gain from a university degree for a young person of type k, and perhaps *the* key object of our analysis:

$$\begin{aligned} ATE(k) \equiv \mathbb{E}[w_1 - w_0 \,|\, k] &= \mathbb{E}[w_1 \,|\, k] - \mathbb{E}[w_0 \,|\, k] \\ &= \mu_1(k) - \mu_0(k). \end{aligned}$$

In appendix C, we show how to aggregate these type-conditional returns to obtain the usual estimands that analysts estimate when considering the returns to education—average treatment effect (ATE), average treatment on the treated (ATT)—within our framework. We also demonstrate some of the biases from which these standard estimation approaches suffer.

The linear factor-model approach that the current state-of-the-art allows one to study the heterogeneity in returns to education, and to compare the contribution of education to

context.

wage dispersion with the contribution of prior ability. One can also study the correlation between cognitive and noncognitive abilities. However, the linearity assumption shuts down any interaction (e.g. complementarities) between the different components of prior ability both on wages directly, and on the returns to education. This approach also requires homoscedasticity: error terms cannot depend on the level of prior ability. We are able to relax this assumption for both the wage and measurement equations.

3 Estimation strategy

Although the non-parametric identification proof detailed in appendix B is constructive and hence suggests a method to operationalise our framework, we prefer an alternative semi-parametric approach via the EM algorithm. In particular, this avoids the necessity of discretizing the measurement and outcome distributions, allowing us to use all the available information in these observations.

Following our approach in Cassagneau-Francis et al. (2021), we assume that the measurements are normally distributed conditional on prior ability, and that log-wages are normally distributed conditional on prior ability and education. Therefore, measurement M_j has probability density function (PDF)

$$f_j(M_j|k) = \phi\left(\frac{M_j - \alpha_j(k)}{\omega_j(k)}\right),$$

where $\phi(\cdot)$ is the standard normal PDF.

Similarly, log-wages, w, are distributed as[18]

$$f(w|k,d) = \frac{1}{\exp w}\phi\left(\frac{w - \mu(k,d)}{\sigma(d)}\right).$$

3.1 EM algorithm

Having made parametric assumptions, we can now use Dempster et al. (1977)'s expectation-maximisation (EM) algorithm to estimate the parameters of the model via maximum likelihood (ML). The computational burden can be further reduced by applying Arcidiacono and Jones (2003)'s sequential-EM algorithm which avoids having to estimate many parameters in one step.

[18] We could allow for heteroscedasticity here, i.e. for the variance of wages to depend on type as well as education, but we found that the algorithm performs better when only the mean of wages varies across types, and not the variance. This may be a consequence of the relatively small samples that we use in the application.

The ML estimator of the parameters, $\Omega = \{\pi(z,d|k), \alpha_j(k), \omega_j(k), \mu(k,d), \sigma(d)\}$, satisfies

$$\hat{\Omega} \equiv \arg\max_{\Omega} \sum_{i=1}^{N} \ln \left(\sum_{k} p_k \ell(\Omega; \boldsymbol{M}_i, w_i, z_i, d_i, k) \right)$$

where $\ell(\Omega; \boldsymbol{M}_i, w_i, z_i, d_i, k) = \pi(z,d\,|\,k) f_m(\boldsymbol{M}\,|\,k) f_w(w\,|\,k)$.

The sum inside the logarithm prohibits sequential estimation of the parameters in Ω.

Arcidiacono and Jones (2003) show the same $\hat{\Omega}$ satisfies

$$\hat{\Omega} \equiv \arg\max_{\Omega} \sum_{i=1}^{N} \sum_{k=1}^{K} p_i(k|\Omega) \ln \ell(\Omega; \boldsymbol{M}_i, w_i, z_i, d_i, k) \tag{4}$$

where

$$p_i(k|\Omega) \equiv \Pr(k|\boldsymbol{M}_i, w_i, z_i, d_i; \hat{\Omega}, \hat{p}) = \frac{\hat{p}_k \ell_i(\hat{\Omega}; \boldsymbol{M}_i, w_i, z_i, d_i, k)}{\sum_{k=1}^{K} \hat{p}_k \ell_i(\hat{\Omega}; \boldsymbol{M}_i, w_i, z_i, d_i, k)}$$

and

$$\hat{p}_k = \frac{1}{N} \sum_{i=1}^{N} p_i(k|\hat{\Omega}).$$

Crucially, the right-hand side of (4) *lends itself to sequential estimation.*

3.2 Bootstrap

As in Cassagneau-Francis et al. (2021), we use a bootstrap procedure to obtain standard errors and confidence intervals for our estimates to account for the random nature of the estimation algorithm. We follow the advice of O'Hagan, Murphy, Scrucca, and Gormley (2019) who recommend using a weighted-likelihood bootstrap (WLBS) to prevent groups from disappearing in any samples. The WLBS involves drawing N positive, non-zero weights from the Dirichlet distribution (which sum to N) ensuring that no observations are completely dropped from any sample. The procedure is computationally intensive as it involves re-estimating our model on 500 such weighted samples. To speed up the procedure and ensure consistent labelling we use the full-sample model estimates as starting values for each bootstrap estimation. We obtain 500 "bootstrapped estimates" for each of our model parameters and can obtain standard errors as standard deviations of these bootstrapped estimates, and confidence intervals from the corresponding quantiles.

4 Application: returns to a UK university degree

We now turn to our application, in which we estimate the returns to a university degree in the UK, as a function of prior ability. To achieve this aim, we apply the framework described in section 2 to data from the 1970 British Cohort Study (BCS 1970), following

the estimation strategy outlined in section 3. We first briefly describe the context of higher education in the UK at the time our data was collected, then discuss the data and the specific variables used to estimate the model. The results follow in section 5.

4.1 Higher education in the UK

The higher education system in the UK has a number of features that make it well suited to studying the returns to *a university degree in general*, as we do in this paper, rather than taking a more granular approach allowing for different types of institutions and degrees. The institutions in the UK were relatively homogenous in what they offer to students. All degree-granting institutions are privately run, and in receipt of government funding.[19] The standard degree offered by a UK university is a three-year bachelor's degree specialising in a single subject, with students generally entering university at age 18 or 19.[20] The student is then awarded a Bachelor of Arts (BA, in arts or humanities subjects) or a Bachelor of Sciences (BSc) in that subject upon graduation.[21] Most universities offer students a wide range of subjects, and have large student bodies, with the largest having nearly 19,000 undergraduate students enrolled in 1994 (HESA, 1996).

A crucial difference across institutions is in their selectivity; universities were able to select students based on their prior attainment at school (as well as at interviews). Therefore, any differences across universities are not likely to be separable from differences in prior ability, a dimension which we explicitly allow returns to vary across in our analysis. Allowing for different types of institution by selectivity would likely not change our analysis. A consequence of the stratifying of universities by ability, however is that we cannot rule out that differences across individuals with different abilities are due to their attendance at different *institutions* and not due to *interaction between ability and higher education.* Separating these effects remains a question for future research.

There were no tuition fees for domestic students during the period when young people in our sample were at university, and there was a system of means-tested grants and loans for "maintenance", i.e. designed to help students cover their living costs (Greenaway and Haynes, 2003). In addition the dropout rate is particularly low, with around 90% of students completing the degree they started in 1989/1990 (Smith and Naylor, 2001). In the UK, leaving home to attend university is a major part of the experience. In the late 1980s and early 1990s when our cohort members were most likely to be at university, over 90% of university students did not live at home (HEFCE, 2009).

[19] There was one university in the UK which did not receive government funding at this time, and was instead run as a charity; the University of Buckingham.

[20] Figure E1 in appendix E contains a detailed timeline of the university application process.

[21] There are a number of other subjects that have their own official title and abbreviation, such as the Bachelor of Laws (LLB), and Bachelor of Engineering (BEng).

In terms of demographics, 46% of women and 49% of men at the "typical age of graduation" in 1996 held an undergraduate degree, with over 13% of both genders also holding a higher degree (OECD, 1998, p. 200). Splitting the population by social class, in 1991 less than 10% of those whose main parent was in the three lowest social classes[22] enrolled in university, while over 35% of children of parents in intermediate roles, and over 50% of children of professional parents enrolled in university (Dearing, 1997). In this paper we abstract from any analysis by family background, focusing only on the role of ability.

4.2 Data

Our data is from the 1970 British Cohort Study (BCS), an ongoing longitudinal cohort study of every person born in the United Kingdom in a week in April 1970. There were 16,568 initial cohort members (CMs), who have been contacted roughly every five years since their birth, with eleven completed "sweeps" to date. The latest sweep is currently underway in 2021 (with the CMs aged 51). In each sweep the CMs (and/or their families) are interviewed about their current circumstances and daily life, with more specific focuses at different stages of their lives. Relevant for the analysis in this paper are measures of cognitive (reading and mathematics tests) and non-cognitive (locus-of-control, self-esteem, mental health) abilities from age 16, and information on qualifications and wages at age 26.

We therefore focus on the fourth sweep,[23] which took place in 1986 (when the CMs were aged 16), and on the fifth sweep which took place a decade later in 1996 (when the CMs were aged 26). These sweeps provide information from just before the decision to attend university, which is generally made at age 17 in the UK, and from when the majority of young people who would attend university have completed their degrees and entered the labour market. We split the sample by gender and estimate the model separately for men and women to enable comparison with previous work on the returns to education during this period, and because there is a significant gender pay gap in the data, both for graduates and non-graduates. Investigating the mechanisms behind the gender pay gap, though interesting and vital, is beyond the scope of this paper.

Table 1 describes the variables that we use to estimate our model. To arrive at our subsample, any CMs who did not respond at either the age 16 or age 26 sweep are dropped. Cohort members are also dropped if they: were not working at age 26; did

[22] In the UK the six social classes are: professional (I), intermediate (II), skilled non-manual (IIIn), skilled manual (IIIm), partly skilled (IV) and unskilled (V).

[23] The fourth sweep was called *Youthscan* at the time, and the data collection was carried out by the *International Centre for Child Studies*. Information was collected from the cohort members themselves, their parents, and their schools (teachers and head teachers). The survey instruments used include questionnaires (both face-to-face and self-completion), medical examinations, diaries, and educational assessments.

Table 1: Description of variables used to estimate our model

Variable	Description
Wage, W	Usual weekly wage in GBP reported by the cohort member if employed at age 26. $w \equiv \log(W)$.
Cognitive score, M^C	Mean standardised score (out of 100) across reading and mathematics tests taken by the cohort members as part of the study at age 16.
Non-cognitive score, M^{NC}	Mean standardised score (zero mean, unit variance) across three measures of "personality": self esteem, locus of control, and the general health questionnaire.†
Desire to leave home, z	Response of cohort member at 16 to the question: "How much do you think [living away from home] will matter to you as an adult?".‡
Education, d	A indicator for whether the cohort member reports holding at least an undergraduate degree at age 26.

Notes: Conti and Heckman (2010) use similar measures from the same dataset to capture non-cognitive abilities and their effects on later health outcomes. †The general health questionnaire (GHQ) is a series of questions designed to predict susceptability to mental health issues. ‡Possible responses: "Matters very much"; "Matters somewhat"; "Doesn't matter".

not take a reading nor a maths test at age 16; did not provide responses to *any* of the non-cognitive measures;[24] are missing information on their highest qualification; or did not respond to the question about leaving home. Finally we trim the sample on wages to keep only those observations with wages between the 1st and 99th percentiles. Table 2 contains summary statistics for the subsample we use for our analysis, pooled and split by education and gender. Wages (W) are increasing in education (d) for both men and women. We denote log-wages by w. The mean graduate ($d = 1$) wage for women is below that of non-graduate men, despite women having similar cognitive test scores and higher non-cognitive measures, supporting our decision to estimate the model separately for men and women. Cognitive (M^C) and non-cognitive (M^{NC}) measures are positively correlated with education for both men and women. Our "instrument" (z), a measure of how strongly an individual wishes to leave home, is positively correlated with holding a degree ($d = 1$) at age 26.

We say "instrument" as z is subject to much weaker exogeneity requirements than a usual instrument. In fact, z need not be an instrument for schooling at all. The conditions z must satisfy are that: (i) it is correlated with type (k) *or* education (d), and (ii) it is independent of wages conditional on k and d. In our application z is an instrument, and therefore we only require that z is independent of wages conditional on type (and education).

[24] We keep individuals for whom we have an incomplete set of cognitive or non-cognitive scores and compute the mean of non-missing scores.

Table 2: Summary statistics for the analysis subsample split by sex and education

Gender:	All	Male			Female		
Education:		All	$D=0$	$D=1$	All	$D=0$	$D=1$
Weekly wage (age 25, GBP, W)							
Mean	239	286	263	334	209	197	241
Std dev.	408	480	481	474	350	388	212
Degree	0.29	0.32	0	1	0.27	0	1
Male	0.40	1	1	1	0	0	0
Ability measures (M)							
Cognitive	57.0	57.8	54.3	65.3	56.5	53.6	64.4
Reading	46.1	46.0	43.0	52.4	46.1	43.6	53.1
Mathematics	40.8	41.4	38.6	48.2	40.5	38.1	47.2
Noncognitive	0.13	0.09	0.01	0.26	0.16	0.10	0.34
Self-esteem	16.5	16.5	16.1	17.3	16.5	16.2	17.1
Locus-of-control	14.3	14.4	14.1	15.0	14.3	14.2	14.5
GHQ†	1.55	1.26	1.23	1.33	1.74	1.60	2.13
Leaving home matters... (z)							
...very much	0.19	0.14	0.13	0.17	0.22	0.20	0.27
...somewhat	0.48	0.47	0.46	0.50	0.48	0.47	0.50
...doesn't matter	0.33	0.39	0.41	0.34	0.30	0.32	0.23
N	1876	745	509	236	1130	827	304

Notes: The values in the table are the mean value of that variable among the population indicated by the column headings, unless otherwise specified. The notation used in the model is in parentheses on the table to highlight which variables in the data correpond to which in the model.

Table E1 in the appendix presents the results of a balancing exercise to provide evidence on the validity of the instrument. This exercise consists of a series of regressions with key (excluded) characteristics as the dependent variable in each regression, and the cognitive and non-cognitive measures, an indicator for females, and our instrument as covariates. The dependent variables are: parental income (in bands); father's (or mother's if father is absent) social class; self-assessed health; whether the young person lived in a city, town, village, or the countryside; and whether the young person is white. These are all observed at age 16. The results are reassuring, with the majority of the coefficients on the instrument not statistically significant, even at the 10% level. Self-assessed health at age 16 is the only exception. Table E2 (also in the appendix) contains the results of a multinomial logit with the young person's region of residence as the dependent, and suggests no evidence of correlation between region and the instrument once we control for prior ability using our measurements. Our balancing exercise suggests the desire to leave home is uncorrelated with other characteristics that might determine wages conditional on prior ability, and is a valid "instrument" for our purposes.

5 Results

This section presents the results of estimating our model on data from the BCS 1970 as we described in the previous sections.[25] We first discuss how to choose the number of types, K, which is also the number of points of support for the distribution of prior ability. We then present results by type using only cognitive ability measurements. This is not our preferred specification. However, non-cognitive measurements are rare, especially in adminstrative datasets, and therefore it is informative to see how our method performs with only cognitive measures. We then estimate our preferred specification which includes measures for both cognitive and non-cognitive prior ability, and finally we compare aggregate results across specifications, and to estimates obtained using more standard estimators.

Throughout this section we label types so that k is increasing in the mean wages of those without a degree, $\mu(k,0)$. By estimating the model separately for men and women, we may estimate a different set of types for men and women, especially if prior ability and wage distributions differ across genders. However, we can compare types within and across

[25] We use the sequential EM algorithm presented in section 3 and appendix D to maximise the sample likelihood (4). We run *kmeans* on $\boldsymbol{M}$ and w to obtain starting values for $\pi(k)$, $\alpha(k)$, and $\mu(k,d)$. We also tried using different starting values and selecting the results with the highest likelihood, but using *kmeans* always produced a likelihood at least as high as the best among the randomly chosen starting values. We use the R programming language to implement our method. The algorithm is relatively fast to converge in our application, taking under one minute on a laptop with a quad-core Intel Core i7-6560U CPU (2.20GHz) processor and 16GB of memory, running Linux (Fedora OS). The variables we use as w, $\boldsymbol{M}$, z, and d are detailed in table 1.

genders using the type-conditional means of ability, $\alpha_C(k)$, and of wages, $\mu(k,d)$.[26]

5.1 Choosing K

The econometrician must set a number of types, K, before estimating the model, and so we estimate the model for K between 2 and 20 and use a range of criteria to select the best choice. What we call the *likelihood criteria*, displayed in figure 1a for the model with only cognitive ability measures estimated on men, are the log-likelihood and the penalised log-likelihood. We are looking for elbows where the slope of the plotted line decreases (all criteria) or maxima (AIC, BIC). The plots in figure 1a suggest picking a value of K less than 7, although the BIC is uninformative. We also study the aggregate results for different K to see if there are any clear patterns, or whether any values of K appear to produce anomalous results. Figure F7 in the appendix is an example of how estimated aggregate returns vary with K.

We also use as a criterion the entropy of the assignment to groups: the uncertainty or "fuzziness" in the assignment, which we assess by studying the distribution of the posterior probabilities, $p_i(k)$. Stronger assignments have clearer modes of the posterior probability distribution at 0 and 1. Figure 1b displays the distribution of posterior probabilities for the same model as figure 1a estimated on women. We would choose $K = 2$ or 3 based on this evidence. The likelihood criteria and posterior probability distributions for other models and samples are shown in appendix F.1. In general the likelihood and entropy criteria suggest we want to select the lowest values of K which capture key patterns in the results.

5.2 Measures of cognitive ability only

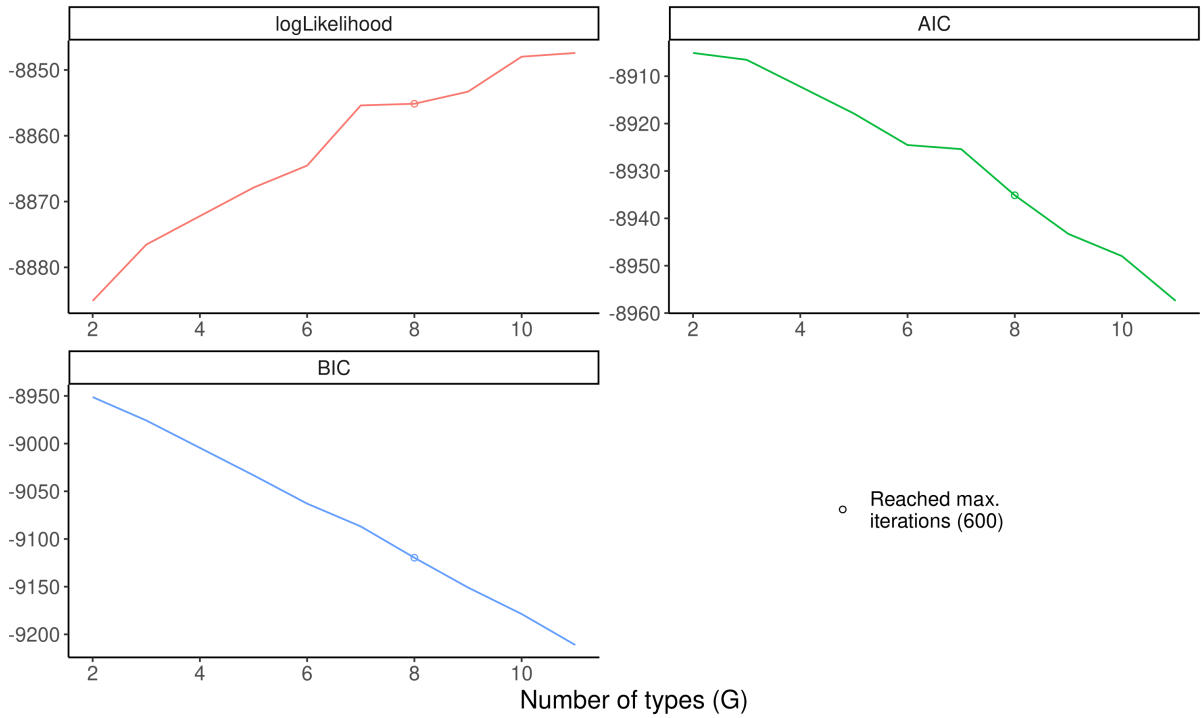
We first present estimates obtained using only cognitive measures of ability. Although this is not our preferred specification, datasets containing only measures of cognitive ability are generally much more widespread than those containing non-cognitive measures (or both), especially in large administrative datasets (so-called "big data"). Therefore, comparing our method using only cognitive measures with our preferred specification is important for understanding the possible limitations of these much larger (in terms of observations), though much less rich (in terms of information) datasets.

Table 3 displays these results for $K = 3$. The results for other values of K, which are broadly similar, are in the appendix (figure F5). Although there is a significant gender wage gap, each of the three types are close in terms of cognitive ability between males and females. For example type 1 men have a mean cognitive score of approximately 45,

[26] Recall that under our model of (noisy) measures and outcomes, the type-conditional means are directly informative of prior ability, although individual measures and outcomes are not.

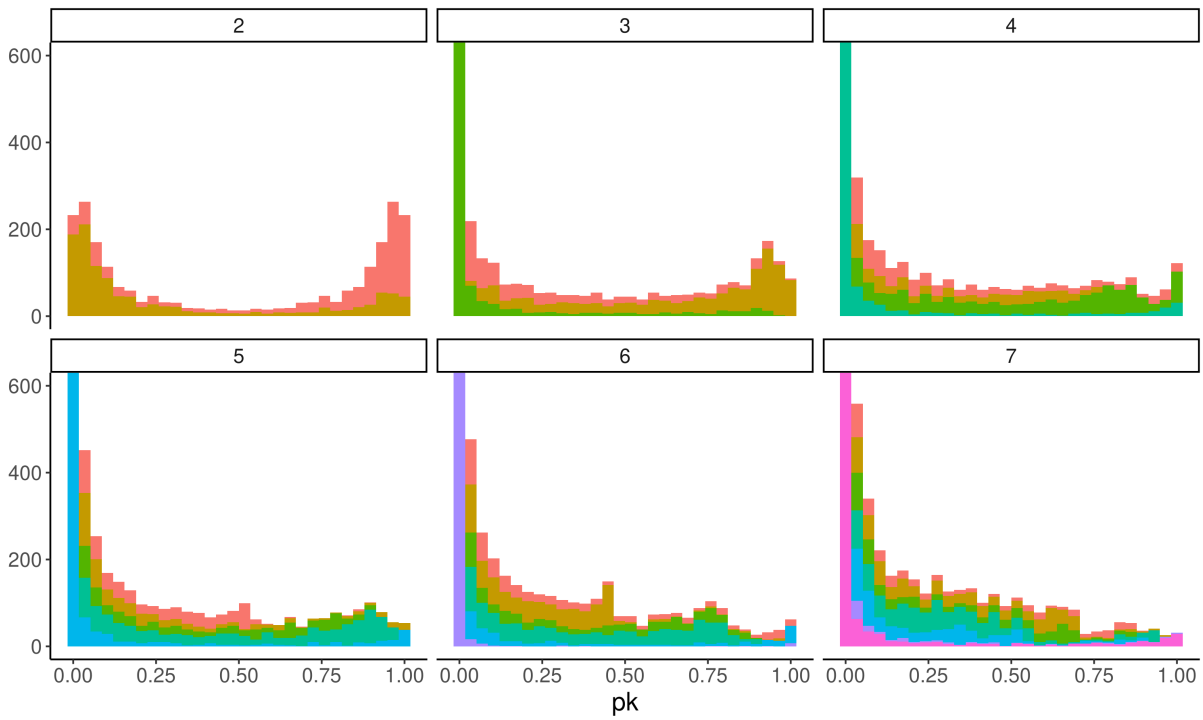
Figure 1: Criteria to select the number of types, K

(a) Likelihood criteria (cognitive measure, males)



Notes: In the top-left panel ("logLikelihood") we plot the loglikelihood (L) of the model against the number of types. In the top-right panel ("AIC") is the negative of the Akaike Information Criterion (AIC), with $AIC = \ln L - 2k$, where k in the number of free parameters. Finally the bottom-left panel ("BIC") plots the negative of the Bayesian Information Criterion (BIC), with $BIC = \ln L - \frac{k}{2} \ln(n)$, with n the number of observations. We are looking for "elbows" (all) and maxima (AIC and BIC). The hollow circles indicate instances in which the algorithm had not converged within 400 iterations.

(b) Distribution of posterior probabilities (cognitive measure, females)



Notes: Each panel represents a different number of types, K. The bars show the distributions of posterior probabilities, $p_i(k)$, coloured according to the value of k. Up to $K = 4$ there are modes at 0 and 1.

Table 3: Results by type ($K = 3$, cognitive measures only)

(a) Male

Type (k)	1		2		3	
Degree	0	1	0	1	0	1
Return to a degree	0.179		0.140		0.239	
Wage (age 25, GBP)						
Mean	205	246	221	254	230	292
Ability measures, $\mathbb{E}[M^{\ell}\|k,d]$						
Cognitive	44.0	43.4	60.8	60.6	77.2	77.4
Non-cognitive	-0.07	0.23	0.06	0.25	0.16	0.28
$\pi(k)$	0.32	0.03	0.32	0.17	0.05	0.12

(b) Female

Type (k)	1		2		3	
Degree (d)	0	1	0	1	0	1
Return to a degree	0.222		0.286		0.188	
Wage (age 25, GBP)						
Mean	147	184	158	210	188	227
Ability measures, $\mathbb{E}[M_j\|\theta_k,d]$						
Cognitive	40.1	36.8	58.3	58.0	77.8	77.6
Non-cognitive	0.03	0.34	0.12	0.35	0.25	0.31
$\pi(k)$	0.21	<0.01	0.49	0.17	0.02	0.09

Notes: The tables in panel (a) and (b) present the key parameter estimates from our model, and their transformations. The returns are in log-differences and are simply the within-type difference between graduate and non-graduate mean log-wages ($\mu(k,1) - \mu(k,0)$). The mean wages at 25 are the type-conditional mean log-wages exponentiated to give weekly wages in GBP, $\exp[\mu(k,d)]$. The cognitive and non-cognitive scores are simply the estimated type-conditional means, and the type proportions are the mean across all men or women of the posterior probabilities, $p_i(k)$, for each type, k.

while type 1 women have a mean cognitive score of 40. Despite having lower wages and similar cognitive scores, the mean non-cognitive scores of the female types are higher than those of the equivalent male type. This highlights the importance of studying men and women separately as they likely face different prices for their abilities on the labour market. Table 3 also splits the type-conditional means by education, d. For the cognitive measure used to estimate the model, the mean functions appear to be independent of education. However, the non-cognitive ability measure *is* correlated with education even *within* types.

Returns to a degree at each level of prior ability are generally higher for women then men, except for those with the highest cognitive ability. The pattern of returns across prior (cognitive) ability also differs across genders. The pattern for men is U-shaped, with those in the middle of the prior ability distribution experiencing lower returns than those with high or low cognitive ability. For women returns are hump-shaped with respect to prior cognitive ability, with middle types enjoying the highest returns to university. These patterns are clearest in figure F5 in the appendix. These non-linearities in returns with respect to cognitive ability highlight the importance of a framework such as ours which does not impose linearity. Given the correlation within types between non-cognitive ability and education, we will withold judgement on whether this pattern of returns to university is robust to the inclusion of both cognitive and noncognitive measures until we present the results of our preferred specification.

This correlation between non-cognitive ability measures and education is apparent in figure F6, where each type is plotted in the space of cognitive and non-cognitive skills. For both men (F6a) and women (F6a) there are large differences within types in terms of non-cognitive ability between graduates (blue) and non-graduates (red). Our method can be considered a *matching estimator*, comparing the outcomes of individuals who graduate from university, with those possessing the same latent characteristics (as captured by their ability measurements and wages) who do not graduate from (or even attend) university. A key takeaway from this analysis is that when we include only a cognitive measure in our model, we are only successful in matching along the cognitive dimension. Therefore, despite the correlation between cognitive and non-cognitive skills it appears to be important to include a measure of non-cognitive ability in the model. This raises questions about the limitations of large administrative datasets, which lack non-cognitive measures, for analyses of the returns to education. Further work is needed to determine whether there are other variables that can be used to proxy for this missing information.

5.3 Measures of cognitive and noncognitive ability

We now present the results of our preferred specification, which includes measures of both cognitive and non-cognitive ability. We estimated the model separately for females and

males in our sample, as these groups appear to face different prices for their skills. We present estimates obtained with $K = 5$ as that is the smallest K which captures key trends and allows us to study variation in both components of ability. We discuss the results for each gender separately. However, before getting to our results we first spend some time explaining the plots in figures 2 and 3 as they are quite particular to our analysis.

The plots in panel (a) of figures 2 and 3 show the same information as those in figure F6 for our model with cognitive and non-cognitive ability measures. The axes represent cognitive (x-axis) and non-cognitive (y-axis) ability measures, so that the location of the circles representing each type-education group reflects their relative prior ability levels. Moving north on the plot represents an increase in the mean non-cognitive ability measure, and moving east represents an increase in the mean cognitive ability measure. The sizes of the circles represent the sizes of the type-education groups, with types labelled in black text, and graduates represented by blue circles and non-graduates by red. Including non-cognitive measures was successful in one sense at least; both cognitive and noncognitive abilities are now independent of education within types. Individuals are well-matched across education groups on both cognitive *and* non-cognitive skills, in contrast to figure F6. Once again types are labelled so that k is increasing in the mean wages of non-graduates, $\mu(k, 0)$.

Turning next to panel (b) of figures 2 and 3, the plots are drawn on the same axes as the plots in panel (a), so the location of the circles again reflects the mean ability measures of that type, now averaged across graduates and non-graduates of each type. However, in panel (b) the sizes of the circles represent mean *log-wages* for each group, with filled circles representing non-graduates, and hollow circles representing graduates. The difference in size between the filled and hollow circles therefore reflects the type-conditional wage return to university, a value which is also labelled on each filled circle in white. These returns are measured in log-wage differences.

5.3.1 Females

Our main results for women are presented in figure 2 and table 4a. Focusing first on the type-education group sizes displayed in figure 2a, university graduation is generally increasing in both cognitive and non-cognitive prior ability. This is reflected by the increasing size of the blue circles relative to the red as we move north-east. However, the relationship is stronger for non-cognitive prior ability. The type-conditional graduation rate,[27] denoted $\Pr(d = 1|k)$ in table 4a, are highest for types 4 and 5, the types with the highest non-cognitive prior ability.

The presentation in figure 2 allows one to easily compare the relative cognitive and non-

[27] This graduation rate is the percentage of all young people (in our analyis) of that type who have graduated from university by age 26.

cognitive abilities for each type, and assess how varying these components affects the returns to university. However, it is difficult to see how to combine the two components into an overall measure of ability. For this we turn to table 4a and remind the reader that types are labelled by non-graduate wage, which is a proxy for overall prior ability, or at least a measure of the relative values of cognitive and non-cognitive abilities on the labour market. Therefore, we can study how returns vary with overall prior ability by studying how they vary across types. The general pattern of returns in table 4a is hump-shaped, with the highest returns to university being achieved by those in the middle of the prior ability distribution. This pattern is most clear in 4a.[28] The returns are still large at around 15 log points for those at the top and bottom of the prior ability distribution, but reach nearly 30 log points in the middle of the distribution. The graduation rates for these types is relatively low, with only 22% of those with the highest return to a degree (type 3) actually graduating from university. The graduation rate improves for type 4, the group with the next highest return, but still less than half of these young women are graduates at age 26.

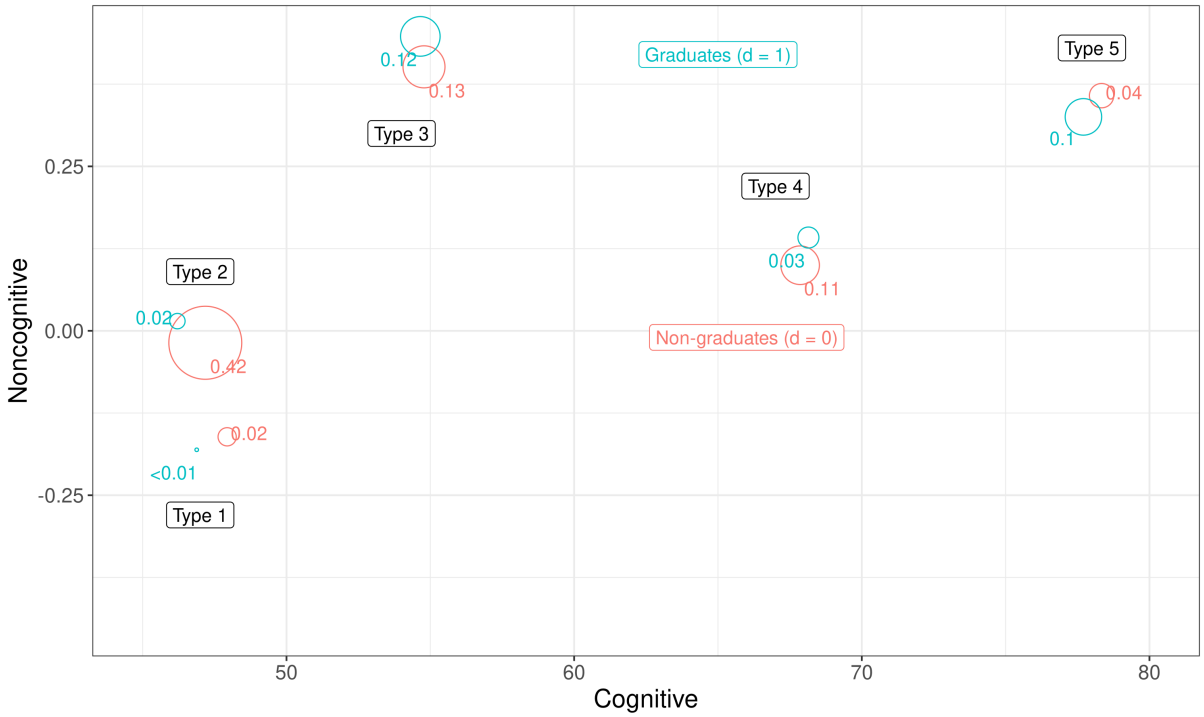
Returning to figure 2b we can see how the different components of ability impact returns. Given the locations of the five types in cognitive-non-cognitive ability "space", it is not immediately obvious how to separate the effects of the two components (the types are not on a "grid", i.e. no two types share the same level of either component). However, some types are similar in one component. For example, moving from type 1 to type 2, involves an increase in non-cognitive ability and a slight decrease in cognitive ability. The returns to university are higher for type 2, suggesting at least at the lower end of the ability distribution increasing non-cognitive ability has a positive effect on returns.

Moving from type 2 to type 3 represents a large increase in cognitive ability and a small increase in non-cognitive ability, and results in both an increase in non-graduate wages and in the returns to a degree. Conversely, moving from type 2 to type 4 represents a small increase in cognitive ability, and a large increase in non-cognitive ability, and again we see increases in both non-graduate wages and the returns to university. This suggests that in this portion of the ability distribution cognitive and non-cognitive abilities are broad substitutes, for both graduates and non-graduates. Finally comparing types 3 and 4 to type 5, the returns to university fall, though this is driven by the relatively high non-graduate wages earned by type 5 women. These young women have the highest cognitive ability, but lower non-cognitive ability than type 4 women, suggesting high cognitive ability women have a comparative advantage as non-graduates (relative to their lower cognitive ability peers), though they still benefit from a university degree.

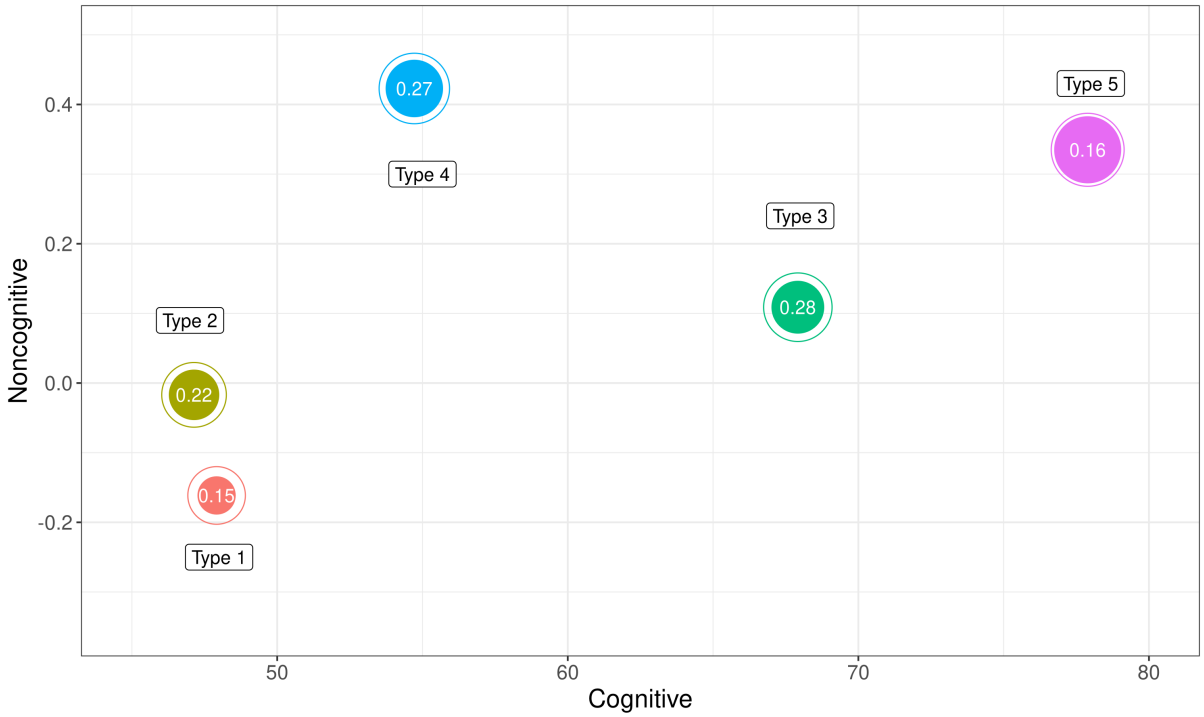
[28] The x-axis in 4a is not type, but type-conditional graduation rate, $\Pr(d = 1|k)$. For women, this results in the same ordering as using $\mu(k, 0)$.

Figure 2: Group sizes, locations and wage premia in cognitive-noncognitive ability space (females, $K = 5$, cognitive and non-cognitive measures)

(a) Group sizes and locations



(b) Wage premia



Notes: Panel (a) display the mean abilities (circle position) and group size (circle sizes and labels) for each type, split education (colour). Blue circles represent graduates and red non-graduates. The size of each type-education group is labelled, along with each type. Panel (b) shows the distribution of wages and wage premia by type, in the space of abilities. The positions of the circles correspond to the cognitive and non-cognitive abilities of that type. The areas of the filled circles are proportional to non-graduate log-wages, and of the hollow circles to graduate log-wages. Then the difference between the areas of filled and unfilled circles is the graduate wage premium, as a difference in log-wages. This wage premium is also labelled in white on each circle.

5.3.2 Males

We turn now to the results for young men, presented in figure 3 and table 4b. The male types follow a broadly similar pattern to those for women, with the bottom two types sharing similarly low levels of cognitive ability, and the key difference between type 1 and type 2 being their non-cognitive ability (see figure 3a). The university graduation rate, $\Pr(d = 1|k)$ is also generally increasing with type, although type 3 men are slightly less likely to graduate from university than their type 2 peers (table 4b). Similar to our findings for women, non-cognitive ability seems important for gaining a university degree.

There is no clear pattern in returns with respect to overall prior ability, measured by non-graduate wages. If we instead order types by graduation rate, as in figure 4b, the returns follow a U-shaped pattern.[29] The U-shape is quite pronounced, with the least and most likely types to graduate from university enjoying returns of over 21 log points, while the "middle" types in terms of graduation rates both having returns below 15 log points. The returns are large for all types at over 11 log points, although the two types with the highest returns enjoy nearly double that of the type with the lowest return.

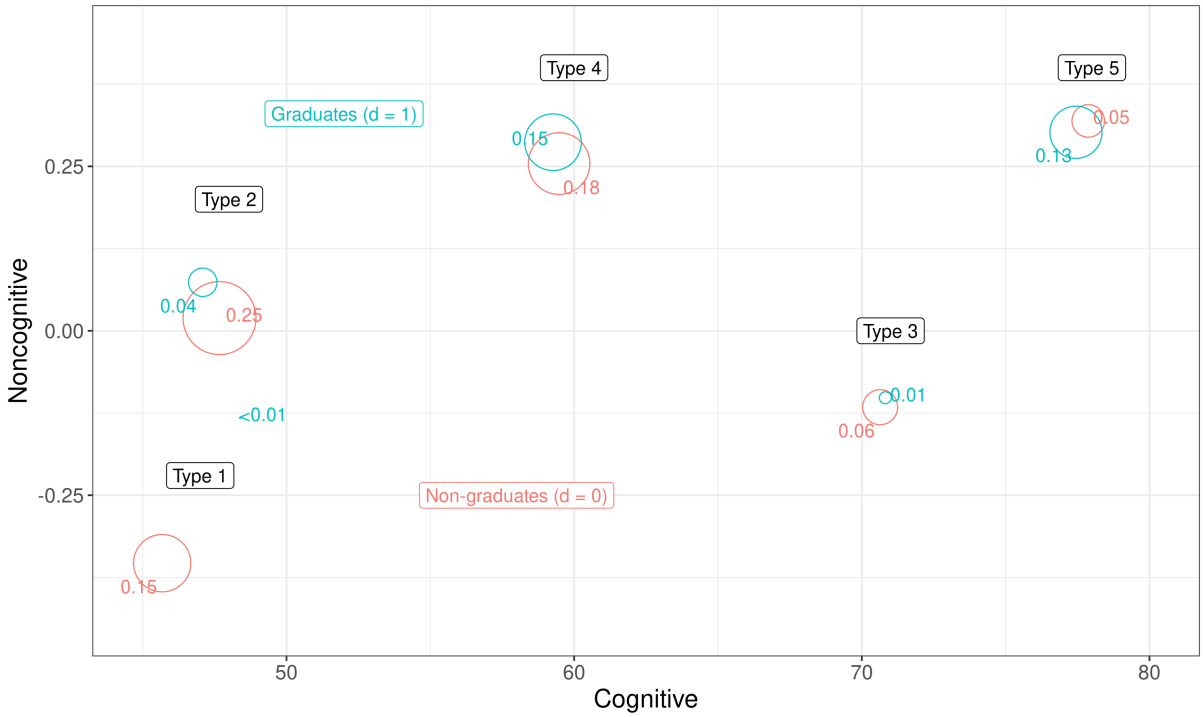
We can also compare types by the cognitive and non-cognitive components of their prior ability, not only their overall prior ability (non-graduate wage). Doing so, we see that types 2 and 4, whose non-cognitive ability is relatively high compared to their cognitive ability, have relatively low returns to a university degree. Moreover, those types with relatively high cognitive ability (types 3 and 5) enjoy the highest returns to university. Although prior non-cognitive skills appear to increase the likelihood of *all* young people gaining a degree, *for men* it is the interaction of prior cognitive skills with higher education that is most valued on the labour market.

Our results suggest that male graduates and non-graduates enter quite different occupations, where prior cognitive skills are better rewarded for graduates, while non-cognitive skills are (relatively) better rewarded for non-graduates. This is despite non-cognitive ability apparently increasing the likelihood of a young person graduating from university. The same cannot be said for women, whose (prior) cognitive and non-cognitive skills appear to be similarly substitutable for both graduates and non-graduates. The analysis in our paper has abstracted from considering occupations separately, including the effects of occupational choice in the "black box" that is the impact of a university degree. However, opening up this black box, including gaining a deeper understanding of how the occupations graduates choose differ from those chosen by non-graduates, will be a key focus for future research.

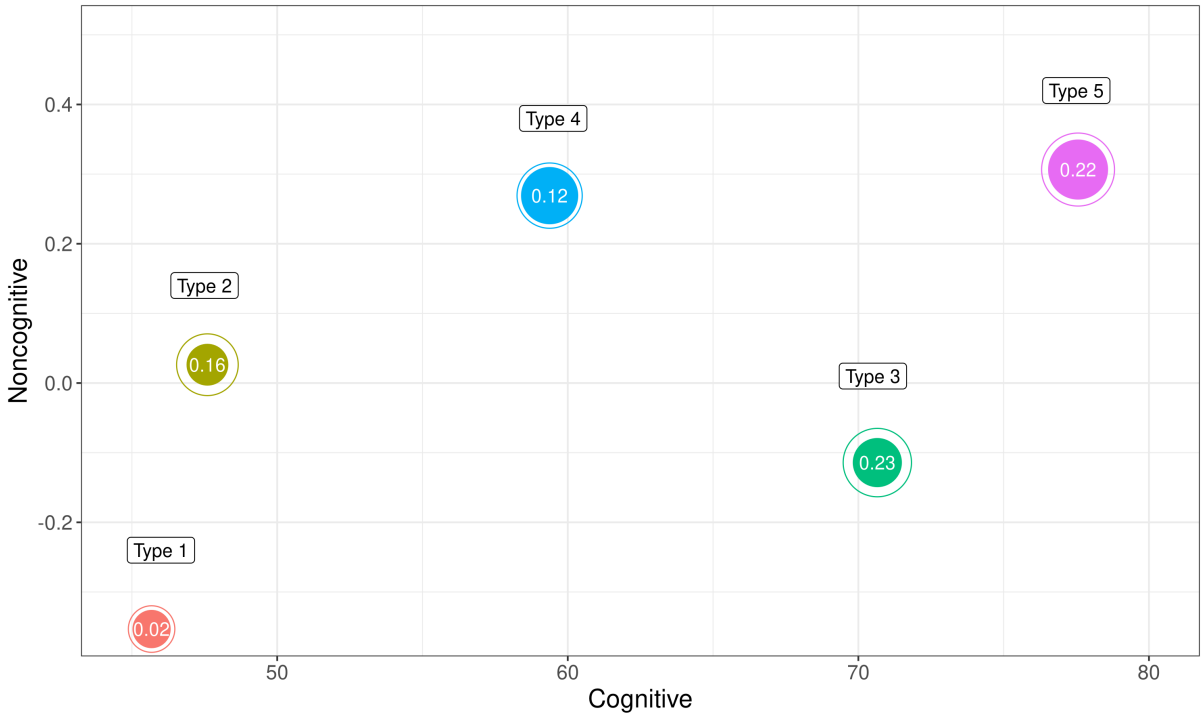
[29] This way of presenting the type-conditional returns is useful as it allows comparison with the marginal treatment effect of Heckman and Vytlacil (2005). Type 1 isomitted from this plot, in part due to its graduation rate of approximately zero.

Figure 3: Group sizes, locations and wage premia in cognitive-noncognitive space (males, $K = 5$, cognitive and non-cognitive measures)

(a) Group sizes and locations



(b) Wage premia



Notes: Panel (a) display the mean abilities (circle position) and group size (circle sizes and labels) for each type, split education (colour). Blue circles represent graduates and red non-graduates. The size of each type-education group is labelled, along with each type. Panel (b) shows the distribution of wages and wage premia by type, in the space of abilities. The positions of the circles correspond to the cognitive and non-cognitive abilities of that type. The areas of the filled circles are proportional to non-graduate log-wages, and of the hollow circles to graduate log-wages. Then the difference between the areas of filled and unfilled circles is the graduate wage premium, as a difference in log-wages. This wage premium is also labelled in white on each circle.

Table 4: Type-conditional mean ability measures and wages
($K = 5$, cognitive and noncognitive measures)

(a) Female

Type (k)	1		2		3		4		5	
Returns	0.150		0.219		0.278		0.267		0.164	
	(0.086)		(0.051)		(0.129)		(0.045)		(0.047)	
$\Pr(d=1\|k)$	0.02		0.04		0.22		0.47		0.70	
Education (d)	0	1	0	1	0	1	0	1	0	1
Wage (age 25, GBP)										
Mean	143	166	151	188	154	204	163	213	192	227
	(12.8)	(11.7)	(2.83)	(9.35)	(9.09)	(36.6)	(4.86)	(7.15)	(5.16)	(8.20)
Ability measures										
Cognitive	47.9	46.9	47.2	46.2	67.9	68.1	54.8	54.7	78.3	77.7
Non-cognitive	-0.16	-0.18	-0.02	0.01	0.10	0.14	0.40	0.45	0.36	0.33
$\pi(k,d)$	0.02	<0.01	0.42	0.02	0.11	0.03	0.14	0.12	0.04	0.10

(b) Male

Type (k)	1		2		3		4		5	
Returns	0.024		0.157		0.228		0.115		0.216	
	(0.051)		(0.108)		(0.167)		(0.067)		(0.079)	
$\Pr(d=1\|k)$	<0.01		0.13		0.09		0.46		0.73	
Education (d)	0	1	0	1	0	1	0	1	0	1
Wage (age 25, GBP)										
Mean	207	212	207	242	213	268	227	255	234	290
	(23.6)	(3.37)	(6.70)	(28.2)	(15.5)	(52.9)	(8.61)	(13.9)	(8.61)	(17.9)
Ability measures										
Cognitive	45.7	48.4	47.7	47.1	70.6	70.8	59.5	59.3	77.9	77.4
Non-cognitive	-0.35	-0.13	0.02	0.07	-0.12	-0.10	0.25	0.29	0.32	0.30
$\pi(k,d)$	0.15	<0.01	0.25	0.04	0.06	0.01	0.18	0.15	0.05	0.13

Notes: The tables in panel (a) and (b) present (transformed) key parameter estimates from our model, with bootstrapped standard errors (500 WLBS replications) in parentheses. The returns are in log-differences and are simply the within-type difference between graduate and non-graduate mean log-wages, $\mu(k,1) - \mu(k,0)$. The mean wages at 25 are the type-conditional mean log-wages exponentiated to give weekly wages in GBP, $\exp[\mu(k,d)]$. The cognitive and non-cognitive scores are simply the estimated type-conditional means, and the type proportions are the mean across all men or women of the posterior probabilities, $p_i(k)$, for each type, k.

Marginal treatment effects. In figure 4 we present the type-conditional wage premiums, plotted against the type-conditional graduation rates. Presenting our results in this fashion makes them comparable to the MTE of Heckman and Vytlacil (2005). The MTE is defined by Heckman and Vytlacil (2005, p. 678) as

$$\Delta^{MTE}(x, u_D) \equiv \mathbb{E}[w_1 - w_0 | X = x, U_D = u_D],$$

where X are observed and and u_D are unobserved components in the decision to attend university. In our setup, a young person's type captures equivalent variation to X and u_D in Heckman and Vytlacil (2005). Therefore, our type-conditional wage premium, $\mathbb{E}[w_1 - w_0|k] = ATE(k)$, is arguably analogous to the MTE. However, the MTE is usually presented ordered by u_D, not by the untreated outcome as we have done. The equivalent of ordering by u_D in our setup is to order types by $\Pr(d = 1|k)$, the type-conditional graduation rate. A strength of our framework is the flexibility in the way we model outcomes and measurements, allowing our MTE analogue to vary equally flexibly.

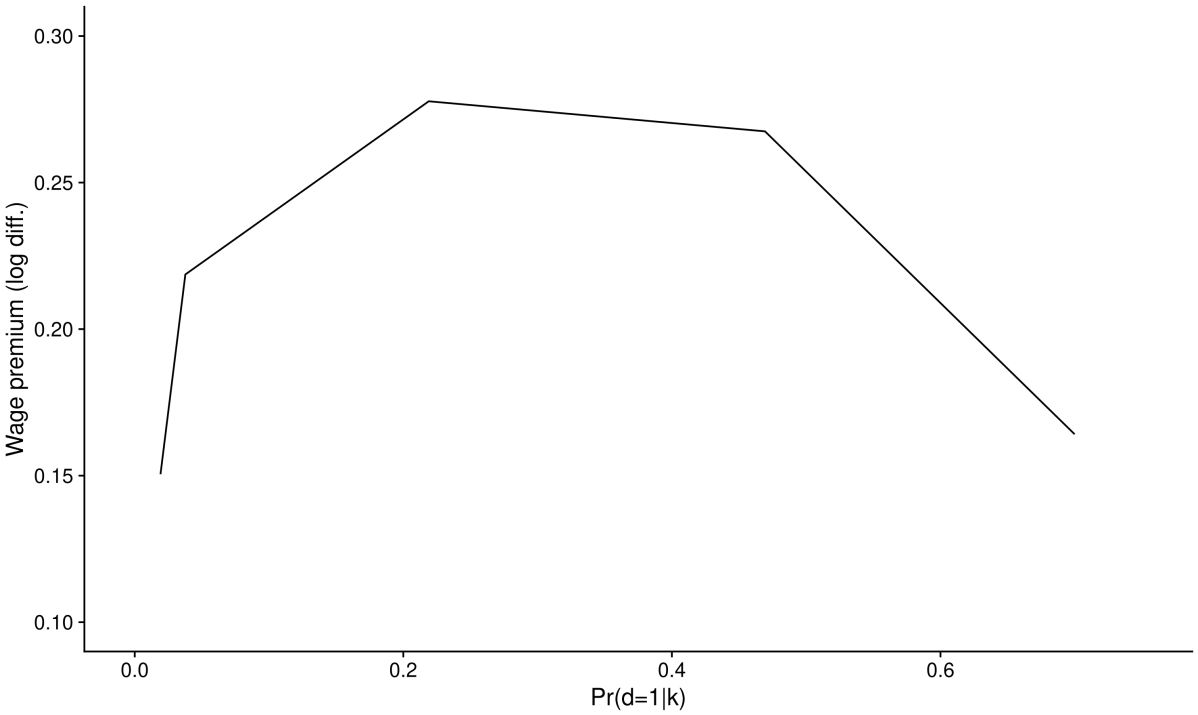
Evidence of non-linearity There is clear evidence of non-linearity in our results. Obtaining the combination of: (i) returns that are increasing in both cognitive and non-cognitive abilities for at least part of their distribution; while (ii) not (monotonically) increasing throughout their distribution, would not have been possible with a linear model. However, due to the correlation between cognitive and non-cognitive skills, and the apparently rather haphazard locations of the types (they do not lie nicely on a grid), determining the source of the non-linearity is difficult. It could be due to non-linearities in the returns to either component — perhaps having higher non-cognitive ability increases returns at the lower end of the distribution, while the opposite is true at the upper end — or it could be due to interactions between the components, or both. Investigating the source of these non-linearities is beyond the scope of this paper.

5.4 Aggregate results

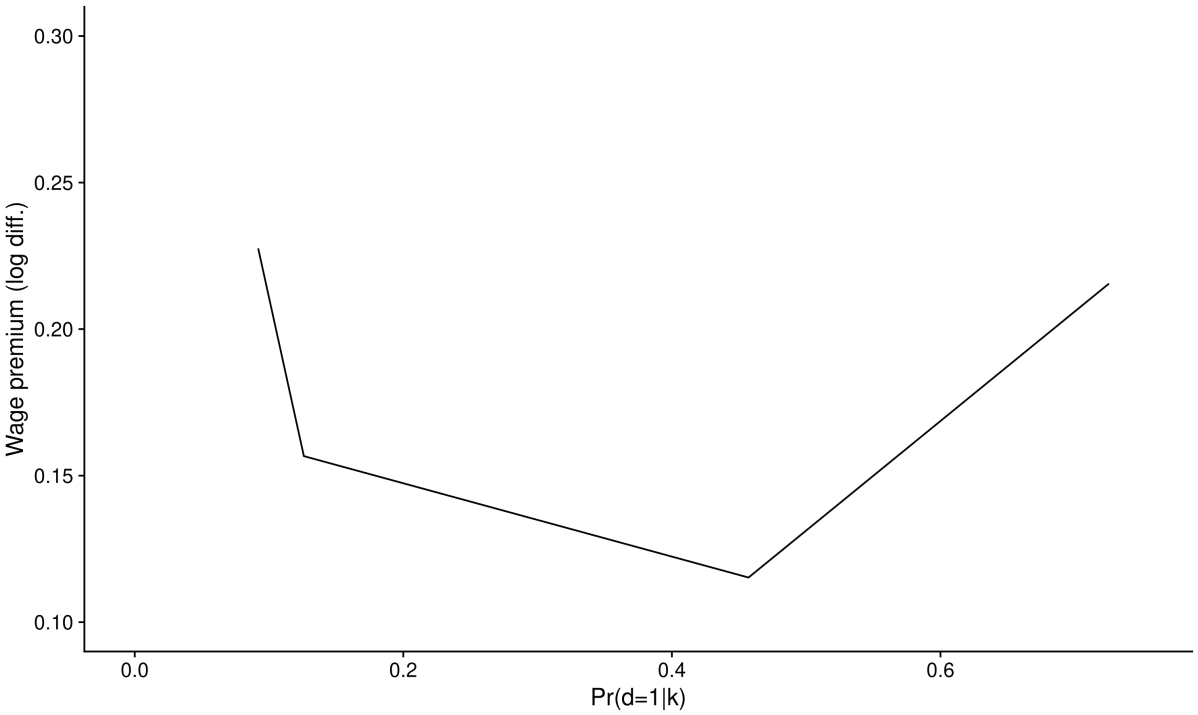
Aggregate results are not a key focus of this paper, which is primarily concerned with estimating heterogeneous returns across people with different levels of prior ability. However, it is still interesting to place our results in the context of previous work, which has generally focused on aggregate returns. In appendix C we show how to aggregate across types to obtain estimates of the average returns across the whole population (ATE) and across only those who chose to attend university (ATT). These estimates are in panel (a) of table 5, along with standard ordinary least squares (OLS), all estimated from our model with $K = 5$. The OLS estimates calculated using our formula (b_{OLS}) are identical to those obtained from an OLS regression of log-wages on education displayed in table F5 (column 1).

Figure 4: Returns by graduation rate

(a) Females



(b) Males



Notes: Panels (a) and (b) plot the type-conditional returns to a degree, $\mathbb{E}[w_1 - w_0|k]$ against the type-conditional graduation rates, $\Pr(d = 1|k)$. The type-conditional return for type-1 males is omitted from panel (b).

Table 5: Aggregate results and comparison with standard estimators ($K = 5$)

(a) Aggregate estimates

Male				Female			
ATE	ATT	b_{OLS}	B_{OLS}	ATE	ATT	b_{OLS}	B_{OLS}
0.137	0.162	0.220	0.058	0.230	0.227	0.325	0.098
(0.060)	(0.053)	(0.042)	(0.031)	(0.038)	(0.042)	(0.033)	(0.026)

(b) Weights in OLS bias (equation 5)

	Male					Female				
Type $k =$	1	2	3	4	5	1	2	3	4	5
Weights	-0.224	-0.252	-0.063	0.212	0.328	-0.030	-0.509	-0.037	0.259	0.316
	(0.197)	(0.158)	(0.183)	(0.204)	(0.210)	(0.097)	(0.096)	(0.018)	(0.090)	(0.086)

Notes: Panel (a) — The values in the table are calculated using the formulas in appendix C. ATE and ATT are average treatment effects and average treatment on the treated. b_{OLS} is the OLS estimator, and our calculated value coincides exactly with the coefficient in a regression of wages on schooling. B_{OLS} is the bias on this estimate versus the "true" ATT. Panel (b) contains the weights used in the formula to calculate the OLS bias. Bootstrapped standard errors (500 WLBS samples) are in parentheses.

Our ATE and ATT estimates are broadly similar to the OLS estimates on our data, and to estimates of other authors on UK data from a similar period.[30] Comparing our model estimates with the OLS and IV estimates using our data raises a number of points worth noting. First, the OLS estimates are broadly similar to the model estimates, though they are slightly biased relative to ATT estimates. In appendix C we derive a formula for the standard OLS estimator of the return to a degree (without controls), b_{OLS}, showing how the OLS estimator is the ATT plus a bias term, B_{OLS}. We reproduce the formula for the bias here.

$$B_{OLS} = \sum_k \underbrace{[\pi(k \mid d = 1) - \pi(k \mid d = 0)]}_{\text{weights}} \mathbb{E}[w_0 \mid k] \tag{5}$$

Panel (b) of table 5 contains the B_{OLS} weights estimated when $K = 5$. Some of these weights are not small, and the relatively small bias on both male and female OLS estimates appears to be due to chance: large positive and negative weights cancel each other out.

5.5 Comparing returns: prior ability versus university

Returning to the results in table 4, we can compare the effects of a low-ability individual graduating from university, with the effects of a (hypothetical) increase in their human capital. The low returns for type 1 of both genders mean they are not a good candidate for such an experiment. However, an interesting comparison involves the wages of a type

[30] Blundell et al. (2000) find wage returns of 17% for men and 37% for women at age 33 using data on a cohort born in 1958.

2 graduate with those of a type 5 (the highest "ability" as measured by wages) non-graduate. For men, type 2 are better off (in wage terms)[31] graduating from university than (hypothetically) increasing their prior ability to the level of the highest ability type (and not attending university). For women, type 2 would earn about the same in either counterfactual. This emphasises how high the wage returns are to a university degree, even for some lower ability young people.

Variance decomposition. The final exercise we perform with the aid of our model is to decompose the variance of wages into three parts:

"within" education groups, due to differences in prior ability;

"between" education groups, due to differences in education;

"unexplained" due to differences in individuals other than education and ability.

Formally, the decomposition is

$$\mathbb{V}(w) = \underbrace{\mathbb{E}[\mathbb{V}(\mathbb{E}[w \mid \theta, d] \mid d)]}_{\text{"within"}} + \underbrace{\mathbb{V}(\mathbb{E}[w \mid d])}_{\text{"between"}} + \underbrace{\mathbb{E}[\mathbb{V}(w \mid \theta, d)]}_{\text{"unexplained"}} \tag{6}$$

which allows us to compare the contributions of prior ability and of the returns to university to wage inequality. The results are in table 6. The majority of the variance in wages is not explained by our model. However, the contribution of the graduate wage premium ("between") to wage inequality is much larger than that of prior ability for both men and women. For women, it is particularly striking, explaining over 23% of the total variance in wages. These findings reinforce the analysis at the end of section 5.3 showing the wage gain from graduating for a low-ability young person (type 2) are equivalent to (hypothetically) being a non-graduate of the highest ability (type 5).

[31] Here we are abstracting from the costs (both pecuniary and non-pecuniary) of graduating from university. These costs, especially the non-pecuniary or "psychic" costs, are likely decreasing in human capital and may be prohibitively high for some low ability young people.

Table 6: Decomposing the variance of log-wages

	Male		Female	
	$\mathbb{V}$	%	$\mathbb{V}$	%
Within (θ)	0.009	3.3	0.014	6.1
Between (d)	0.024	9.2	0.053	23.1
Unexplained	0.229	87.5	0.162	70.7
Total	0.262	100	0.229	100

6 Conclusion

In this paper we have presented a framework designed to separately estimate the effects of ability and higher education on wages. We incorporate insights from the literatures on both human capital formation and the importance of non-cognitive as well as cognitive skills. Our model therefore resembles those in the literature on human capital, but one of our key innovations is a novel nonparametric identification strategy. Although we are not the first to show non-parametric identification, our approach requires fewer measurements of prior ability than the current leading approaches in the literature. We are also the first to take an important next step, estimating our model without imposing linearity in wages nor in measurements.

We demonstrate our method in an application on data from a longitudinal cohort study in the UK. We show that a measure of cognitive ability is not sufficient to fully capture variation in (multidimensional) ability across individuals before attending university, despite strong positive correlation between cognitive and non-cognitive abilities. When we estimate our preferred specification, which includes measures of both cognitive and non-cognitive abilities, we find important non-linearities in the effects of prior ability on wages, and on the returns to a university degree. The returns to university are also shown to be more important than the returns to prior ability: a low ability young person is better off as a low-ability graduate than they would be if instead they were to increase their ability to match their highest-ability peers. The large impact of university on wages across the ability distribution leads to another of our main results: the contribution of the graduate wage premium to inequality is three to four times larger than the contribution of ability.

The implications of our findings are somewhat unsettling. We are not the first to highlight the contribution of (non-universal) higher education to wage inequality (Autor, 2014). According to our results, sending everyone (or no-one) to university would be preferable to the current situation. Moreover, given we find that returns are generally increasing

in prior ability, no higher education is preferred (in terms of inequality) to universal higher education. This is clearly not a policy that many would (or should) support. However, finding ways to mitigate the contributions of higher education to inequality while preserving its many other benefits, both to the individual and society, is vital.

There are also a number of caveats to mention regarding the work in this paper. First, the framework used in this paper (and its sibling in Cassagneau-Francis et al. 2021) is new and needs to be studied in more detail. Second, this is a static, statistical analysis and so does not allow for any equilibrium considerations. Also, in our application we only consider the short term effects of higher education. In future work we plan to expand the model to allow for earnings over a longer period. Another important task is to study how the returns to higher education have evolved over recent decades. Estimating our framework on a more recent cohort would allow such an analysis.

References

Altonji, J. G. and T. A. Dunn (1996): "The Effects of Family Characteristics on the Return to Education," *The Review of Economics and Statistics*, 78, 692–704, publisher: The MIT Press.

Arcidiacono, P. and J. B. Jones (2003): "Finite Mixture Distributions, Sequential Likelihood and the EM Algorithm," *Econometrica*, 71, 933–946, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00431.

Ashenfelter, O. and C. Rouse (1998): "Income, Schooling, and Ability: Evidence from a New Sample of Identical Twins," *The Quarterly Journal of Economics*, 113, 253–284, publisher: Oxford University Press.

Autor, D. H. (2014): "Skills, education, and the rise of earnings inequality among the "other 99 percent"," *Science*, 344, 843–851.

Barrow, L. and C. E. Rouse (2005): "Do Returns to Schooling Differ by Race and Ethnicity?" *The American Economic Review*, 95, 83–87, publisher: American Economic Association.

Becker, G. (1964): *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*, National Bureau of Economic Research, Inc.

Blundell, R., L. Dearden, A. Goodman, and H. Reed (2000): "The Returns to Higher Education in Britain: Evidence from a British Cohort," *The Economic Journal*, 110, F82–F99.

Blundell, R., L. Dearden, and B. Sianesi (2005): "Evaluating the effect of education on earnings: models, methods and results from the National Child Development

Survey," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168, 473–512.

Bonhomme, S. (2021): "Teams: Heterogeneity, Sorting, and Complementarity," *SSRN Electronic Journal*.

Bonhomme, S., K. Jochmans, and J.-M. Robin (2016a): "Estimating multivariate latent-structure models," *The Annals of Statistics*, 44, 540–563, publisher: Institute of Mathematical Statistics.

——— (2016b): "Non-parametric estimation of finite mixtures from repeated measurements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 211–229.

——— (2017): "Nonparametric estimation of non-exchangeable latent-variable models," *Journal of Econometrics*, 201, 237–248.

Bonhomme, S., T. Lamadon, and E. Manresa (2019): "A Distributional Framework for Matched Employer Employee Data," *Econometrica*, 87, 699–739, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA15722.

——— (2022): "Discretizing Unobserved Heterogeneity," *Econometrica*, 90, 625–643.

Bonhomme, S. and E. Manresa (2015): "Grouped Patterns of Heterogeneity in Panel Data: Grouped Patterns of Heterogeneity," *Econometrica*, 83, 1147–1184.

Bonhomme, S. and J.-M. Robin (2009): "Consistent noisy independent component analysis," *Journal of Econometrics*, 149, 12–25.

——— (2010): "Generalized Non-Parametric Deconvolution with an Application to Earnings Dynamics," *The Review of Economic Studies*, 77, 491–533, publisher: [Oxford University Press, Review of Economic Studies, Ltd.].

Bowles, S., H. Gintis, and M. Osborne (2001): "The Determinants of Earnings: A Behavioral Approach," *Journal of Economic Literature*, 39, 1137–1176.

Britton, J., L. Dearden, and B. Waltmann (2021a): "The returns to undergraduate degrees by socio-economic group and ethnicity," IFS Report.

Britton, J., L. van der Erve, C. Belfield, L. Dearden, A. Vignoles, M. Dickson, Y. Zhu, I. Walker, L. Sibieta, and F. Buscha (2021b): "How much does degree choice matter?" Tech. rep., The IFS.

Card, D. (1999): "The Causal Effect of Education on Earnings," in *Handbook of Labor Economics*, Elsevier, vol. 3, 1801–1863.

Carneiro, P., K. T. Hansen, and J. J. Heckman (2003): "Estimating Distributions

of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice*," *International Economic Review*, 44, 361–422.

Carneiro, P., J. J. Heckman, and E. Vytlacil (2010): "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," *Econometrica*, 78, 377–394.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101, 2754–2781.

Cassagneau-Francis, O., R. J. Gary-Bobo, J. Pernaudet, and J.-M. Robin (2021): "A Nonparametric Finite Mixture Approach to Difference-in-Difference Estimation, with an Application to Professional Training and Wages," *Unpublished manuscript*.

Cawley, J., J. Heckman, and E. Vytlacil (2001): "Three observations on wages and measured cognitive ability," *Labour Economics*, 8, 419–442.

Conti, G. and J. J. Heckman (2010): "Understanding the Early Origins of the Education-Health Gradient: A Framework That Can Also Be Applied to Analyze Gene-Environment Interactions," *Perspectives on Psychological Science*, 5, 585–605.

Cunha, F. and J. Heckman (2007a): "The Technology of Skill Formation," *American Economic Review*, 97, 31–47.

Cunha, F. and J. J. Heckman (2007b): "Identifying and Estimating the Distributions of Ex Post and Ex Ante Returns to Schooling," *Labour Economics*, 14, 870–893.

——— (2008): "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Journal of Human Resources*, 43, 738–782.

Cunha, F., J. J. Heckman, L. Lochner, and D. V. Masterov (2006): "Interpreting the Evidence on Life Cycle Skill Formation," in *Handbook of the Economics of Education*, Elsevier, vol. 1, 697–812.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78, 883–931, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA6551.

Dearing, R. (1997): "Higher Education in the learning society," Tech. rep.

Delaney, L., C. Harmon, and M. Ryan (2013): "The role of noncognitive traits in undergraduate study behaviours," *Economics of Education Review*, 32, 181–195.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): "Maximum likelihood

from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, 1–38.

FRENCH, E. AND C. TABER (2011): "Chapter 6 - Identification of Models of the Labor Market," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card, Elsevier, vol. 4, 537–617.

GARY-BOBO, R. J., M. GOUSSÉ, AND J.-M. ROBIN (2016): "Grade retention and unobserved heterogeneity," *Quantitative Economics*, 7, 781–820, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE524.

GREENAWAY, D. AND M. HAYNES (2003): "Funding Higher Education in the UK: The Role of Fees and Loans," *The Economic Journal*, 113, F150–F166, publisher: [Royal Economic Society, Wiley].

GRILICHES, Z. (1977): "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica*, 45, 1–22, publisher: [Wiley, Econometric Society].

HECKMAN, J. AND B. SINGER (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52, 271.

HECKMAN, J., J. STIXRUD, AND S. URZUA (2006): "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 24, 411–482.

HECKMAN, J. J. AND E. VYTLACIL (2005): "Structural equations, treatment effects, and econometric policy evaluation," *Econometrica*, 73, 669–738.

HECKMAN, J. J. AND E. J. VYTLACIL (1999): "Local instrumental variables and latent variable models for identifying and bounding treatment effects," *Proceedings of the National Academy of Sciences*, 96, 4730–4734.

——— (2007): "Chapter 71 Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6, 4875–5143.

HEFCE (2009): "Patterns in higher education: living at home," .

HESA (1996): "Students in Higher Education Institutions 1994/95," .

IMBENS, G. W. AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475, publisher: [Wiley, Econometric Society].

Jacob, B. A. (2002): "Where the boys aren't: non-cognitive skills, returns to school and the gender gap in higher education," *Economics of Education Review*, 21, 589–598.

Kane, T. J. and C. E. Rouse (1995): "Labor-Market Returns to Two- and Four-Year College," *The American Economic Review*, 85, 600–614, publisher: American Economic Association.

Kotlarski, I. (1967): "On characterizing the gamma and the normal distribution." *Pacific Journal of Mathematics*, 20, 69–76, publisher: Pacific Journal of Mathematics, A Non-profit Corporation.

Mincer, J. (1958): "Investment in human capital and personal income distribution," *Journal of political economy*, 66, 281–302.

——— (1974): *Schooling, experience, and earnings*, no. 2 in Human behavior and social institutions, New York: National Bureau of Economic Research; distributed by Columbia University Press.

OECD, ed. (1998): *Education at a glance OECD indicators*, Paris: OECD.

O'Hagan, A., T. B. Murphy, L. Scrucca, and I. C. Gormley (2019): "Investigation of Parameter Uncertainty in Clustering Using a Gaussian Mixture Model Via Jackknife, Bootstrap and Weighted Likelihood Bootstrap," *arXiv:1510.00551 [stat]*, arXiv: 1510.00551.

Reiersol, O. (1950): "Identifiability of a Linear Relation between Variables Which Are Subject to Error," *Econometrica*, 18, 375–389, publisher: [Wiley, Econometric Society].

Smith, J. P. and R. A. Naylor (2001): "Dropping out of university: A statistical analysis of the probability of withdrawal for UK university students," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164, 389–405.

Taber, C. R. (2001): "The Rising College Premium in the Eighties: Return to College or Return to Unobserved Ability?" *Review of Economic Studies*, 68, 665–691.

Todd, P. E. and K. I. Wolpin (2003): "On the Specification and Estimation of the Production Function for Cognitive Achievement," *The Economic Journal*, 113, F3–F33.

Todd, P. E. and W. Zhang (2020): "A dynamic model of personality, schooling, and occupational choice," *Quantitative Economics*, 11, 231–275.

A Linear model

A common assumption made to help identify and estimate models like the one above is that both wages and measurements are linear in their components. Then,

$$w_d = \mu_d^0 + \mu_d^C \theta^C + \mu_d^N \theta^N + \varepsilon_d \tag{7}$$
$$M_\ell = \gamma_\ell^0 + \gamma_\ell^C \theta^C + \gamma_\ell^N \theta^N + \varepsilon_\ell. \tag{8}$$

This is the approach taken by Cunha and Heckman (2007a,b), henceforth CH, for example.

Assumption (Independent errors)**.** *The error terms, ε_d and ε_ℓ are independent of θ and each other, and have means equal to zero.*

Under this assumption we obtain the *classical measurement error model*, and OLS estimates using M to proxy θ as in the following equation,

$$w_d = \delta_d^0 + \boldsymbol{M}'\delta_d + \eta_d \tag{9}$$

where $\delta = (\delta^1, ..., \delta^L)$, are biased as

$$\begin{aligned}\mathbb{E}[\eta_d M_\ell] &= \mathbb{E}[(\epsilon_d - \boldsymbol{\varepsilon}'\delta_d)(\gamma_\ell^0 + \gamma_\ell^C \theta^C + \gamma_\ell^N \theta^N + \varepsilon_\ell)] \\ &= \delta_d \mathbb{E}[\varepsilon_\ell^2] \neq 0,\end{aligned}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_L)$, and the first equality is obtained by combining equations (7) and (8) to match equation (9) and equating the error terms. The second equality follows from assumption **??**. Therefore, we cannot recover θ nor any of the parameters in equations (7) and (8) via OLS. However, models with classical measurement error are well studied in economics and statistics. When w and M are *not* jointly normal, Reiersol (1950) shows that the parameters in equations (7) and (8) are identified, up to some normalisations. More recent work has shown how to identify the distribution of θ and the error terms using a theorem due to Kotlarski (1967).[32] Bonhomme and Robin (2009, 2010) generalise these results to allow for the non-parametric identification and estimation of such factor models.

B Nonparametric identification proof

We first state the necessary conditions under which our model is identified.

Assumption 1 (Measurements and wages)**.** *Measurements, wages and z are independent conditional on type and education.*[33]

[32] See Carneiro et al. (2003) for more details.

[33] Measurements need not be independent of each other even conditional on type.

Assumption 2 (No empty cells). $\pi(k,0,d) \neq 0$ *for all* d *and for all* k.

Assumption 2 ensures that for at least one value of the instrument, arbitrarily set to zero, young people of all endowments of prior ability have positive probability of both attending and not attending university.

Assumption 3 (Linear independence). $\left[f_M(\boldsymbol{M}|1) \quad \cdots \quad f_M(\boldsymbol{M}|k) \quad \cdots \quad f_M(\boldsymbol{M}|K)\right]$ *and, for all* d, $\left[f_w(w|1,d) \quad \cdots \quad f_w(w|k,d) \quad \cdots \quad f_w(w|K,d)\right]$ *are linearly independent systems.*

Assumption 3 means we cannot identify any points of support in the distribution of human capital for which the associated conditional measurement and / or wage distribution can be formed by a linear combination of the distributions corresponding to other points of support. This is analogous to the rank condition in ordinary least squares.

Assumption 4 (First stage). $\dfrac{\pi(k,1,d)}{\pi(k,0,d)} \neq \dfrac{\pi(k',1,d)}{\pi(k',0,d)}$ *for all* d, *for all* $k \neq k'$.

Assumption 4 ensures that exposure to the instrument leads to different sized shifts in university attendance for individuals with different levels of prior ability. It is analogous to the rank condition in IV estimation.

Assumption 5 (Measurements independent of education). *For all types* k *and all measurements* M_ℓ, $f_\ell(M^\ell|k,d) = f_\ell(M^\ell|k)$.

Assumption 5 allows us to label groups consistently across education levels. There are other assumptions that we could make to achieve the same aim. However, it seems reasonable to assume that conditional on ability, our measurements of ability are independent of later education.

Assumption 6 (Discrete wages and measurements). *The distributions of wages and measurements have discrete support.*

Assumption 6 is not strictly necessary but it is a relatively innocuous assumption that greatly simplifies the exposition. We could discretise continuous distributions by projecting them onto some functional basis, i.e. $(\boldsymbol{M},w) \mapsto p(z,d,\boldsymbol{M},w)$, and it is straightforward to adapt the proof.

Theorem 1 (Identification). *Under assumptions 1-6 plus the conditional exclusion restriction on the instrument,* $\pi(k,z,d)$, $f_M(\boldsymbol{M}|k) = \prod_\ell f_\ell(M_\ell|k)$, *and* $f_w(w|k,d)$ *are non-parameterically identified.*

Proof. The proof contains three steps.

Step 1: Constructing matrices

The probability of observing an individual with variables $(z_i, d_i, \boldsymbol{m}_i, w_i)$ in our model writes

$$p(z_i, d_i, \boldsymbol{m}_i, w_i) = \sum_k \pi(k, z_i, d_i) f_m(\boldsymbol{m}_i|k) f_w(w_i|k). \tag{10}$$

Under assumption 6, $f_m(\cdot|k)$ and $f_w(\cdot|k)$ are probability mass functions (pmfs), which we place in matrices along with the observable probabilities, $p(z_i, d_i, \boldsymbol{m}_i, w_i)$, and the joint type-instrument-treatment probabilities, $\pi(k, z_i, d_i)$. The matrices (one per z, d value pair) containing the observed data shares,

$$\underset{n_{\boldsymbol{m}} \times n_w}{P(z,d)} \equiv \left[p(z, d, \boldsymbol{m}, w)\right]_{\boldsymbol{m} \times w}$$

are indexed by measurement down their rows, and by wages across their columns. It has dimension $n_{\boldsymbol{m}} \times n_w$. The matrix of type-instrument-treatment probabilities for each k, z pair,

$$\underset{K \times K}{D(z,d)} \equiv \text{diag}\left[\pi(k, z, d)\right]_{k \times k}$$

is a diagonal matrix with dimension K, containing the type-instrument-treatment probabilities, $\pi(k, z, d)$, on its diagonal. Finally, there are the two matrices containing the measurement and wage pmfs

$$\underset{n_{\boldsymbol{m}} \times K}{F_1} \equiv \left[f_m(\boldsymbol{m}|k)\right]_{\boldsymbol{m} \times k} \quad \text{and} \quad \underset{n_w \times K}{F_2} \equiv \left[f_w(w|k, d)\right]_{w \times k},$$

where F_1 is indexed by measurement down its rows, F_2 by wages down its rows, and both matrices by type across their columns. Then, $n_{\boldsymbol{m}}$ is the number of points of support in the (discrete) measurement distribution and the number of rows in F_1, and n_w is the number of points of support in the (discrete) wage distribution, and the number of rows in F_2. Both matrices have K columns.

For a given z, d, we can write equation (10) in matrix form

$$P(z,d) = F_1 D(z,d) F_2(d)^\top.$$

Step 2: Identifying F_1, $D(z,d)$, and $F_2(d)$

For a given d (which we omit to simplify the notation) the following matrices, corresponding to the different values of z,[34] share the same algebraic structure[35]

$$P(0) = F_1 D(0) F_2^\top$$
$$P(1) = F_1 D(1) F_2^\top$$

as F_1 and F_2 are independent of z. Assumption 2 ensures $D(0)$ and $D(1)$ are invertible, and by assumption 3, the matrices F_1 and F_2 have full column rank. Therefore $P(0)$ has rank K and admits a singular value decomposition (SVD)

$$P(0) = U\Sigma V^\top,$$

where U and V are rank-$n_{\boldsymbol{m}}$ and rank-n_w unitary matrices. We can partition $U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}$ and $V = \begin{bmatrix} V_1 & V_2 \end{bmatrix}$ so that

$$P(0) = U_1 \Sigma_1 V_1^\top, \tag{11}$$

where Σ_1 contains the K non-zero singular values of $P(0)$ on its diagonal. U_1 is $n_{\boldsymbol{m}} \times K$, V_1 is $n_w \times K$, and Σ_1 is $K \times K$. From the components of the SVD in equation (11), we can construct the matrices $W_1 = \Sigma_1^{-\frac{1}{2}} U_1^\top$ and $W_2 = \Sigma_1^{-\frac{1}{2}} V_1^\top$. Then, applying W_1 and W_2 to $P(0)$ as follows, we obtain Q and Q^{-1}

$$\begin{aligned} W_1 P(0) W_2^\top &= \Sigma_1^{-\frac{1}{2}} U_1^\top U_1 \Sigma_1 V_1^\top V_1 \Sigma_1^{-\frac{1}{2}} = I_K \\ &= \underbrace{W_1 F_1}_{Q} \underbrace{D(0) F_2^\top W_2^\top}_{Q^{-1}} = QQ^{-1} = I_K. \end{aligned}$$

We can similarly apply W_1 and W_2 to $P(1) := P(1,d)$, to obtain

$$W_1 P(1) W_2^\top = \underbrace{W_1 F_1}_{Q} D(1) \underbrace{F_2^\top W_2^\top}_{D(0)^{-1}Q^{-1}} = QD(1)D(0)^{-1}Q^{-1}.$$

The non-zero (diagonal) entries of

$$D(1)D(0)^{-1} = \operatorname{diag}\left[\frac{\pi(k,1,d)}{\pi(k,0,d)}\right]_K$$

are the (unique) eigenvalues of $W_1 P(1) W_2^\top$, which is derived using only the observable matrices P(1) and P(0). Q contains eigenvectors of $W_1 P(1) W_2^\top$, though these eigenvectors are only determined up to a multiplicative constant.

[34] This example is for a binary z, but the proof is easily extended to any discrete, finite z.

[35] By this we mean they can be decomposed into a trio of matrices, where the first and third matrices are identical (F_1 and $F_2^\top$) and the middle matrix is diagonal.

To pin down these eigenvectors, recall that U is unitary and hence

$$U_2^\top P(0) = U_2^\top U_1 \Sigma_1 V_1^\top = 0_{(n_{\boldsymbol{m}}-K)\times n_w}$$

which implies

$$U_2^\top P(0) = U_2^\top F_1 D(0) F_2^\top = 0_{(n_{\boldsymbol{m}}-K)\times n_w}. \tag{12}$$

By assumptions 2 and 3, $D(0)F_2^\top$ has full row rank, so equation (12) implies $U_2^\top F_1 = 0_{(n_{\boldsymbol{m}}-K)\times n_w}$. Now define $\widehat{Q}$ as some matrix of eigenvectors of $W_1 P(1) W_2^\top$, such that there is a diagonal matrix Δ which satisfies $\widehat{Q} = Q\Delta = \Sigma_1^{-\frac{1}{2}} U_1^\top F_1 \Delta$, and $\Sigma_1^{\frac{1}{2}}\widehat{Q} = U_1^\top F_1 \Delta$. Therefore

$$\begin{pmatrix} \Sigma_1^{\frac{1}{2}}\widehat{Q} \\ 0_{(n_{\boldsymbol{m}}-K)\times n_w} \end{pmatrix} = U^\top F_1 \Delta, \tag{13}$$

and

$$U_1 \Sigma_1^{\frac{1}{2}} \widehat{Q} = U \begin{pmatrix} \Sigma_1^{\frac{1}{2}}\widehat{Q} \\ 0_{(n_{\boldsymbol{m}}-K)\times n_w} \end{pmatrix} = UU^\top F_1 \Delta = F_1 \Delta. \tag{14}$$

Then $F_1\Delta = U_1\Sigma_1^{\frac{1}{2}}\widehat{Q}$ is identified, and also we have that $F_1 = U_1\Sigma_1^{\frac{1}{2}}\widehat{Q}\Delta^{-1}$. Noticing the rows of F_1 must sum to one (as each column is a probability distribution), we can find the non-zero (diagonal) elements of Δ using

$$(\Delta_1, ..., \Delta_K) = (1, ..., 1) U_1 \Sigma_1^{\frac{1}{2}} \widehat{Q}, \tag{15}$$

which identifies Δ and hence F_1.

Finally, $\Delta\widehat{Q}^{-1} = Q^{-1} = D(0)F_2^\top V_1 \Sigma_1^{\frac{1}{2}}$, and hence $Q^{-1}\Sigma^{\frac{1}{2}} = D(0)F_2^\top V_1$. V is an unitary matrix, so $P(0)V_2 = 0$, using that $V_1^\top V_2 = 0$, and $F_1 D(0)$ has rank K, implying $F_2^\top V_2 = 0_{n_w\times(n_{\boldsymbol{m}}-K)}$. Then, following similar steps to before,

$$\begin{aligned} Q^{-1}\Sigma^{\frac{1}{2}} V_1^\top &= \begin{pmatrix} D(0)F_2^\top V_1 & 0_{n_w\times(n_{\boldsymbol{m}}-K)} \end{pmatrix} \begin{pmatrix} V_1^\top \\ V_2^\top \end{pmatrix} \\ &= \begin{pmatrix} D(0)F_2^\top V_1 & D(0)F_2^\top V_2 \end{pmatrix} V^\top \\ &= D(0)F_2^\top V V^\top \\ &= D(0)F_2^\top. \end{aligned} \tag{16}$$

The rows of F_2 also sum to one, so $D(0)$ and hence F_2 are identified from equation (16), following a similar argument the one used above to identify Δ and F_1. And $D(1)$ is known now we know $D(0)$ and $D(1)D(0)^{-1}$.

Step 3: Correct labels across d.

We need to ensure that the labels on types are consistent across treatments (i.e. values of d). We use that F_1 is independent of d to ensure that each type is labelled the same across all treatments.

□

C Treatment effects in our framework

As in Cassagneau-Francis et al. (2021), here we show that we can identify some of the usual treatment effect (TE) estimands and their associated biases using our framework.

Average treatment effect. We can aggregate over types to obtain the ATE in (3).

$$ATE \equiv \mathbb{E}[w_1 - w_0] = \sum_k \pi(k) ATE(k)$$

where $\pi(k) = \sum_{z,d} \pi(k,z,d)$, the proportion of young people of type k.

Average treatment on the treated. We can also aggregate over those who attend university within each type to obtain the ATT.

$$ATT \equiv \mathbb{E}[w_1 - w_0 \,|\, d=1] = \sum_k \pi(k \,|\, d=1) ATE(k)$$

where $\pi(k \,|\, d=1) = \frac{\sum_z \pi(k,z,1)}{\sum_{k,z} \pi(k,z,1)}$, the proportion of individuals of type k among those who attend university.

Ordinary least squares (OLS). We can also calculate the OLS estimator, b_{OLS}, within our framework, and decompose this estimand into an ATT term and an "OLS bias" term, B_{OLS}.

$$\begin{aligned} b_{OLS} &= \frac{\text{Cov}(w,d)}{\mathbb{V}(d)} = \mathbb{E}[w_1 \,|\, d=1] - \mathbb{E}[w_0 \,|\, d=0] \\ &= \sum_k \pi(k \,|\, d=1)\, \mathbb{E}[w_1 \,|\, k] - \pi(k \,|\, d=0)\, \mathbb{E}[w_0 \,|\, k] \\ &= ATT + B_{OLS} \end{aligned}$$

where

$$B_{OLS} = \sum_k [\pi(k \,|\, d=1) - \pi(k \,|\, d=0)]\, \mathbb{E}[w_0 \,|\, k]$$

The OLS bias disappears only if (i) $\pi(k\,|\,d=1)=\pi(k\,|\,d=0)$ for all values of k; or (ii) $\mathbb{E}[w_0\,|\,k]=\mathbb{E}[w_0\,|\,k']$ for all $k\neq k'$. The first equality is unlikely to hold in our application as those with higher prior are more likely to attend university, and hence the proportion of those with high prior is likely to be larger among graduates. This is the issue of selection on ability that was mentioned earlier. The second equality is also unlikely to hold, as young people with higher prior ability are generally more productive workers and can hence command a higher wage.

IV and LATE. Finally, we can perform a similar exercise to decompose the standard (two-stage least squares) IV estimator for a binary instrument, into a LATE term which corresponds to Imbens and Angrist (1994)'s local average treatment effect, and an "IV bias" term, B_{IV}.

The two-stage least squares estimator of the effect of university on wages (without controls) is

$$b_{IV}=\frac{\text{Cov}(w,z)}{\text{Cov}(d,z)}=\frac{\mathbb{E}[w|z=1]-\mathbb{E}[w|z=0]}{\mathbb{E}[d|z=1]-\mathbb{E}[d|z=0]}$$

In our framework, the denominator of b_{IV} is

$$\mathbb{E}[d|z=1]-\mathbb{E}[d|z=0]=\sum_k\left[\pi(k,d=1\,|\,z=1)-\pi(k,d=1\,|\,z=0)\right].$$

The numerator has a more interesting decomposition, such that we can write

$$b_{IV}=LATE+B_{IV}$$

where

$$LATE=\sum_k\frac{\pi(k,d=1|z=1)-\pi(k,d=1|z=0)}{\sum_k\left[\pi(k,d=1|z=1)-\pi(k,d=1|z=0)\right]}ATE(k), \tag{17}$$

and

$$B_{IV}=\sum_k\frac{\pi(k|z=1)-\pi(k|z=0)}{\sum_k\left[\pi(k,d=1|z=1)-\pi(k,d=1|z=0)\right]}\mathbb{E}[w_0\,|\,k],$$

with

$$\pi(k,d|z)=\frac{\pi(k,z,d)}{\sum_{k,d}\pi(k,z,d)}\quad\text{and}\quad\pi(k|z)=\sum_d\pi(k,d|z).$$

Therefore the $LATE$ estimator of Imbens and Angrist (1994) is a weighted average of type-specific ATEs in our framework, with weights corresponding to the proportion of *compliers*[36] with that level of prior ability. Note the similarity between our decomposition of the $LATE$ in equation (17) and Heckman and Vytlacil (1999, 2005, 2007)'s marginal

[36] Those who are induced into attending university by the instrument.

treatment effect (MTE):[37]

$$LATE = \frac{\int_{\varphi(0)}^{\varphi(1)} \Delta^{MTE}(\nu)\,\mathrm{d}\nu}{\varphi(1)-\varphi(0)} \tag{18}$$

where

$$\Delta^{MTE}(\nu) = \mathbb{E}\left[w_1 - w_0 \,|\, \nu_i = \nu\right],$$

and $\varphi(z)$ and ν are the observed and unobserved components of the non-pecuniary cost of attending university. The formula in (18) is a weighted average of the returns to university over those induced to attend by the instrument, though this average is over the distribution of (unspecified) unobserved costs, rather than (imperfectly observed) prior ability. Our framework is more flexible in one sense as it allows correlation between outcomes (w_d) and unobserved costs through latent types.

D EM algorithm details

The EM algorithm iterates back and forth over the following two steps:

E-step.

The E-step updates the posterior type probabilities, $p_i(k|\Omega)$:

$$p_i\left(k|\hat{\Omega}^{(s)}\right) \equiv \frac{\hat{p}_k^{(s)}\ell\left(\hat{\Omega}^{(s)};\boldsymbol{M}_i,w_i,z_i,d_i,k\right)}{\sum_{k=1}^{K}\hat{p}_k^{(s)}\ell\left(\hat{\Omega}^{(s)};\boldsymbol{M}_i,w_i,z_i,d_i,k\right)}, \tag{19}$$

where $\Omega = \{\pi(z,d|k), \alpha_j(k), \omega_j(k), \mu(k,d), \sigma(d)\}$, over all values such that $z \in \{0,1\}$, $d \in \{0,1\}$, $k \in \{1,...,K\}$, and $j \in \{C,N\}$.

M-step.

While in the M-step we update the components of Ω in the $(s+1)$-th iteration, using the estimates from the s-th iteration.

- Update $\alpha_j(k), \omega_j(k)$.
    1. Update $\alpha_j(k)$ as the weighted mean test score, using posterior probabilities as weights (for each type)

$$\alpha_j(k)^{(s+1)} \equiv \frac{\sum_i p_i\left(k|\hat{\Omega}^{(s)}\right) M_{ji}}{\sum_i p_i\left(k|\hat{\Omega}^{(s)}\right)} \tag{20}$$

    2. Then $\omega_j(k)$ is updated as the weighted root-mean-square error, using posteriors

[37] This formula is adapted from the presentation in French and Taber (2011)'s excellent survey on the identification of models of the labour market.

as weights

$$\omega_j(k)^{(s+1)} \equiv \sqrt{\frac{1}{N}\sum_{i=1}^{N} p_i\left(k|\hat{\Omega}^{(s)}\right)\left(M_{ji} - \alpha_j(k)^{(s+1)}\right)^2} \tag{21}$$

- Update $\mu(k,d)$, $\sigma(k,d)$.

  1. Again use weighted means, with weights $p_i(k|\hat{\Omega}^{(s)})$ to update $\mu(k,d)$:

$$\mu(k,d)^{(s+1)} \equiv \frac{\sum_{i:d_i=d} p_i\left(k|\hat{\Omega}^{(s)}\right) w_i}{\sum_{i:d_i=d} p_i\left(k|\hat{\Omega}^{(s)}\right)} \tag{22}$$

  2. And use the updated $\mu(k,d)$ to update $\sigma(d)$:

$$\sigma(d)^{(s+1)} \equiv \sqrt{\frac{\sum_k \sum_{i:d_i=d} p_i\left(k|\hat{\Omega}^{(s)}\right)\left(w_i - \mu_d^{(s+1)}(k)\right)^2}{\sum_k \sum_{i:d_i=d} p_i\left(k|\hat{\Omega}^{(s)}\right)}} \tag{23}$$

- Finally, we sum posterior probabilities by k, z, and d to obtain $\pi(k,z,d)$,

$$\pi(k,z,d)^{(s+1)} \equiv \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I(z,d)} p_i(k|\hat{\Omega}^{(s)}), \tag{24}$$

  where $I(z,d) = \{i : z_i = z, d_i = d\}$.

Iterations stop when the algorithm converges, i.e. when the increase in likelihood between iterations is below a threshold:

$$\mathcal{L}(\Omega^{(s)}; \boldsymbol{M}, w, z, d) - \mathcal{L}(\Omega^{(s-1)}; \boldsymbol{M}, w, z, d) < \delta, \tag{25}$$

for some $\delta > 0$ chosen by the econometrician.

E Context and data

Figure E1: Timeline of educational decisions

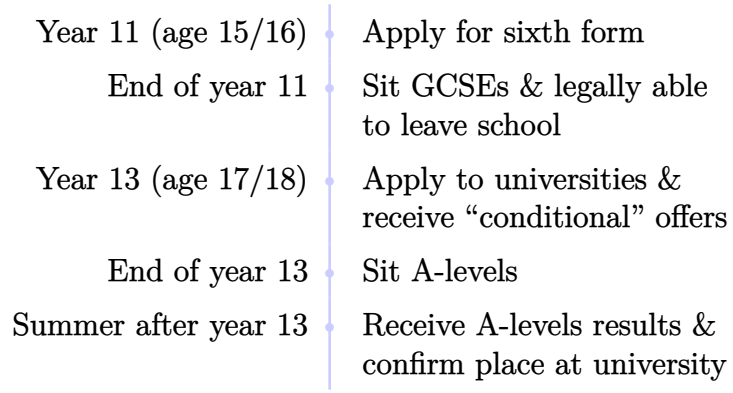


Table E1: Balancing checks for instrument validity

	Dependent variable:				
	Parental				
	Income	Social class	Health	Urban	White
	(1)	(2)	(3)	(4)	(5)
Female	−0.223**	−0.188*	0.245***	0.084	−0.008
	(0.097)	(0.100)	(0.094)	(0.092)	(0.008)
Cognitive	0.016***	0.024***	−0.005	0.004	0.002***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.0003)
Non-cognitive	0.236***	0.250***	−0.180**	0.047	0.004
	(0.083)	(0.088)	(0.080)	(0.078)	(0.007)
Leaving home...					
matters somewhat	−0.070	−0.047	0.236*	−0.133	0.006
	(0.128)	(0.131)	(0.123)	(0.120)	(0.011)
doesn't matter	−0.185	−0.126	0.069	−0.180	0.009
	(0.137)	(0.139)	(0.131)	(0.129)	(0.012)
Observations	1,398	1,418	1,870	1,822	1,816
R^2					0.023
Adjusted R^2					0.020
Residual Std. Error					0.170
F Statistic					8.452***

Notes: *p<0.1; **p<0.05; ***p<0.01
In columns (1–4) the dependent variables are categorical and ordered logit was used to regress the dependent variable on the covariates, using the `polr` function from the R package MASS. In column (5), as the dependent variable is binary we estimate a linear probability model using function `lm` from the R stats package.

Table E2: Balancing checks for instrument validity: regions

	Dependent variable:									
	North West	Yorkshire	East Mids	West Mids	East	London	South East	South West	Wales	Scotland
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Female	−0.179	−0.437*	−0.498*	−0.444*	−0.535**	−0.515*	−0.444*	0.017	−0.512*	−0.201
	(0.252)	(0.258)	(0.272)	(0.264)	(0.254)	(0.290)	(0.243)	(0.281)	(0.294)	(0.259)
Cognitive	0.009	−0.017**	−0.019**	−0.011	−0.002	−0.017*	0.005	−0.0002	−0.012	−0.004
	(0.008)	(0.008)	(0.009)	(0.009)	(0.008)	(0.009)	(0.008)	(0.009)	(0.010)	(0.008)
noncogScore	0.106	0.440**	0.404*	0.215	0.251	0.056	0.212	0.228	0.036	0.199
	(0.207)	(0.214)	(0.227)	(0.219)	(0.211)	(0.241)	(0.200)	(0.227)	(0.245)	(0.212)
Leaving home...										
matters somewhat	0.293	0.038	0.378	0.380	0.537	0.138	0.350	0.159	0.071	0.275
	(0.314)	(0.319)	(0.352)	(0.334)	(0.329)	(0.370)	(0.299)	(0.339)	(0.368)	(0.316)
doesn't matter	0.262	0.017	0.314	0.090	0.413	0.115	−0.091	0.026	−0.092	−0.177
	(0.328)	(0.332)	(0.366)	(0.353)	(0.343)	(0.385)	(0.319)	(0.358)	(0.388)	(0.339)
Akaike Inf. Crit.	8,754.993	8,754.993	8,754.993	8,754.993	8,754.993	8,754.993	8,754.993	8,754.993	8,754.993	8,754.993

Notes: *p<0.1; **p<0.05; ***p<0.01. We use the function `multinom` from the R package nnet to perform the multinomial logit used to estimate the coefficients in this table as region is an unordered categorical variable.

F Results

F.1 Choosing K

F.2 Single cognitive measure

F.3 Cognitive and non-cognitive measures

Table F1: Distribution parameter estimates (male, cognitive and non-cognitive measures)

	$K=5$									
Type$(k)=$	1		2		3		4		5	
$\alpha_C(k)$	45.7		47.6		70.7		59.4		77.6	
$\omega_C(k)$	8.99		13.6		4.78		9.06		6.60	
$\alpha_N(k)$	−0.35		0.03		−0.11		0.27		0.31	
$\omega_N(k)$	0.71		0.37		0.60		0.46		0.50	
$d=$	0	1	0	1	0	1	0	1	0	1
$\mu(k,d)$	5.33	5.35	5.33	5.49	5.36	5.59	5.43	5.54	5.46	5.67
$\sigma(d)$	0.46	0.52	0.46	0.52	0.46	0.52	0.46	0.52	0.46	0.52

Table F2: Distribution parameter estimates (female, cognitive and non-cognitive measures)

	$K=5$									
Type$(k)=$	1		2		3		4		5	
$\alpha_C(k)$	47.9		47.1		67.9		54.7		77.9	
$\omega_C(k)$	9.25		12.6		4.86		7.88		6.77	
$\alpha_N(k)$	−0.16		−0.02		0.11		0.42		0.33	
$\omega_N(k)$	1.21		0.52		0.54		0.56		0.51	
$d=$	0	1	0	1	0	1	0	1	0	1
$\mu(k,d)$	4.96	5.11	5.02	5.23	5.04	5.32	5.09	5.36	5.26	5.42
$\sigma(d)$	0.52	0.42	0.52	0.42	0.52	0.42	0.52	0.42	0.52	0.42

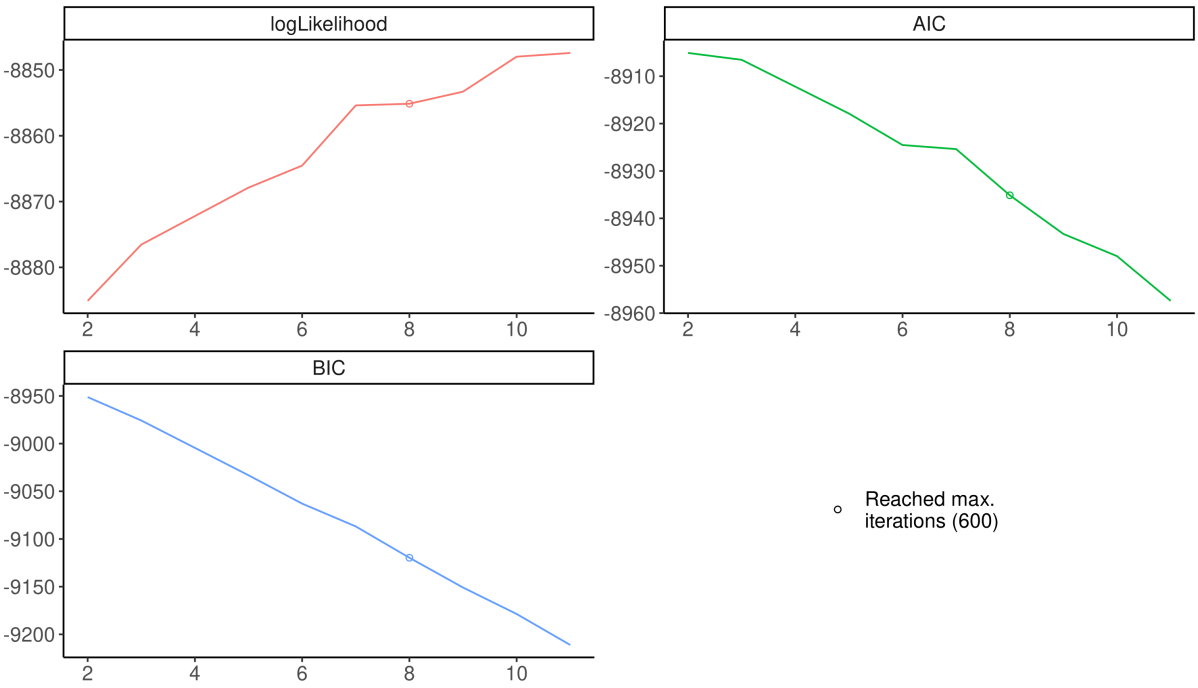
F.4 OLS estimates

In table F5, we present ordinary least squares (OLS, columns 1–4) and two-stage least squares (2SLS, columns 5 and 6) estimates of the returns to a university degree, across a range of specifications. The baseline regression equation is

$$w_i = \beta_0 + \mu_d d_i + \gamma_C M_i^C + \gamma_N M_i^N + X_i'\beta_1 + \varepsilon_i$$

Figure F1: Likelihood criteria: single cognitive measure

(a) Male



(b) Female

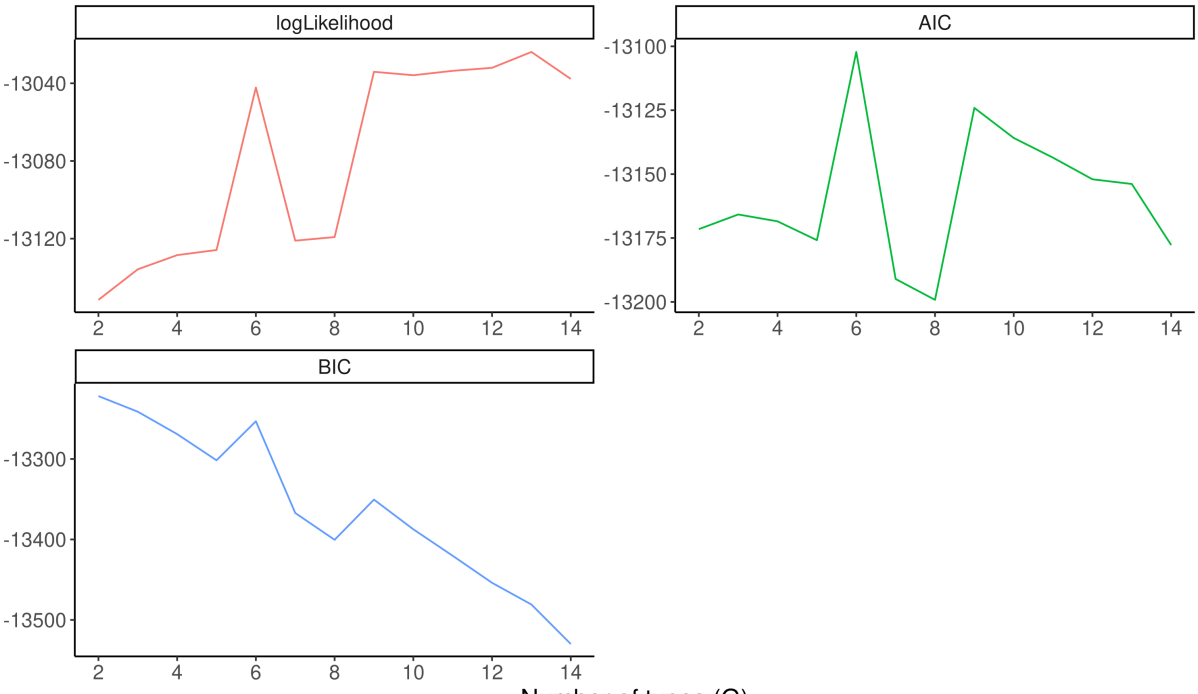
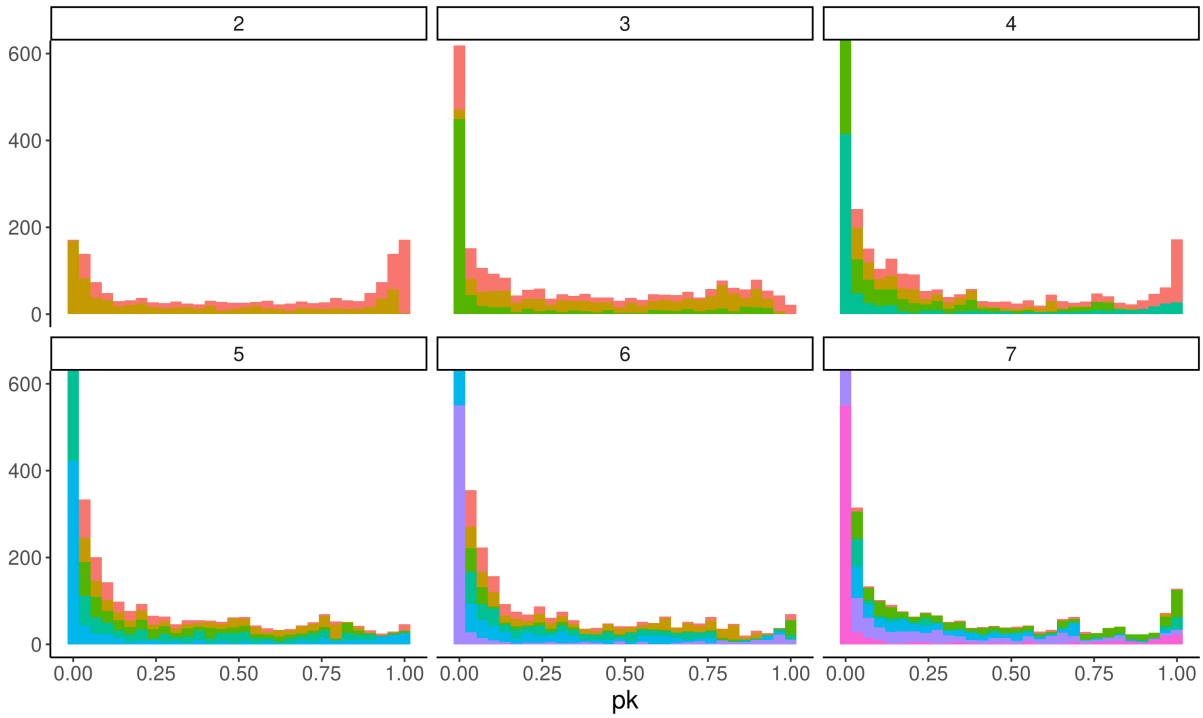


Figure F2: Posterior probabilities: single cognitive measure

(a) Male



(b) Female

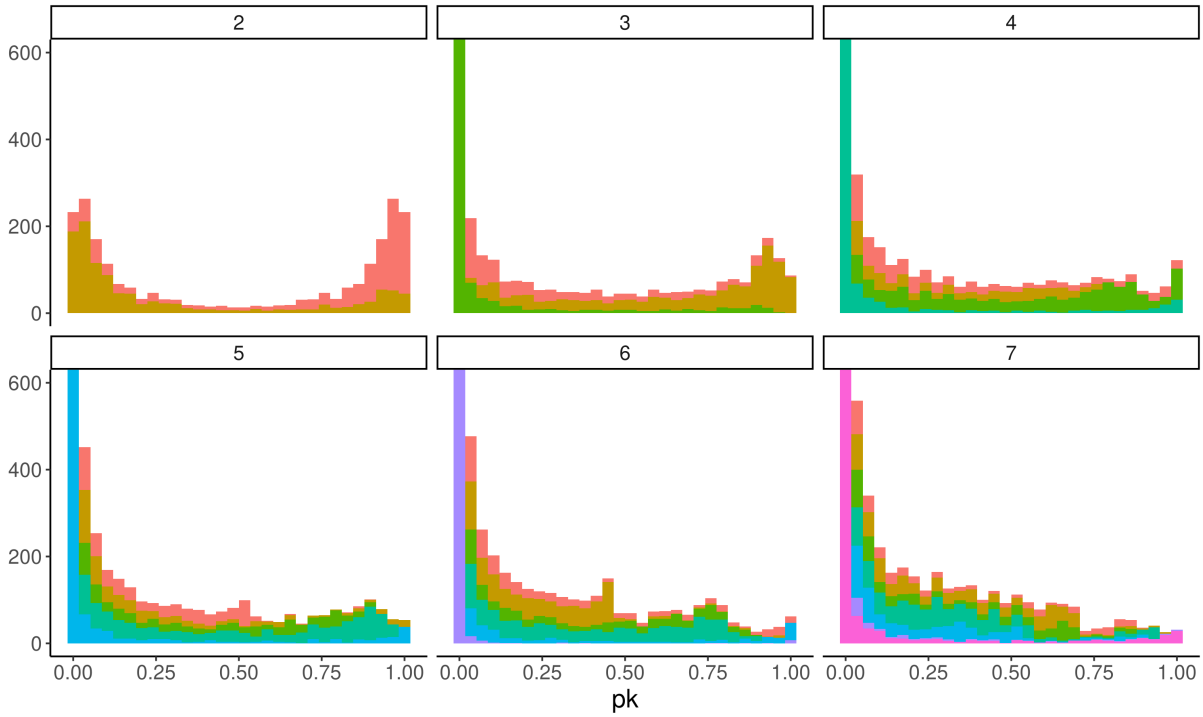
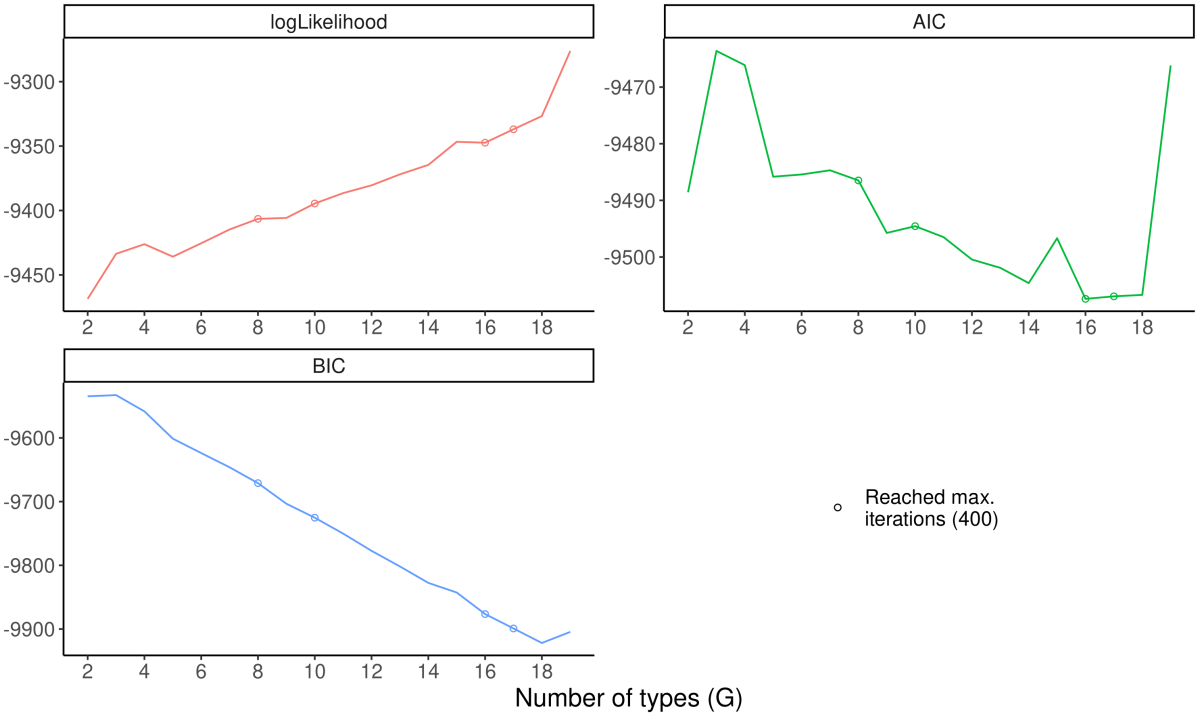


Figure F3: Likelihood criteria: cognitive and non-cognitive measures

(a) Male



(b) Female

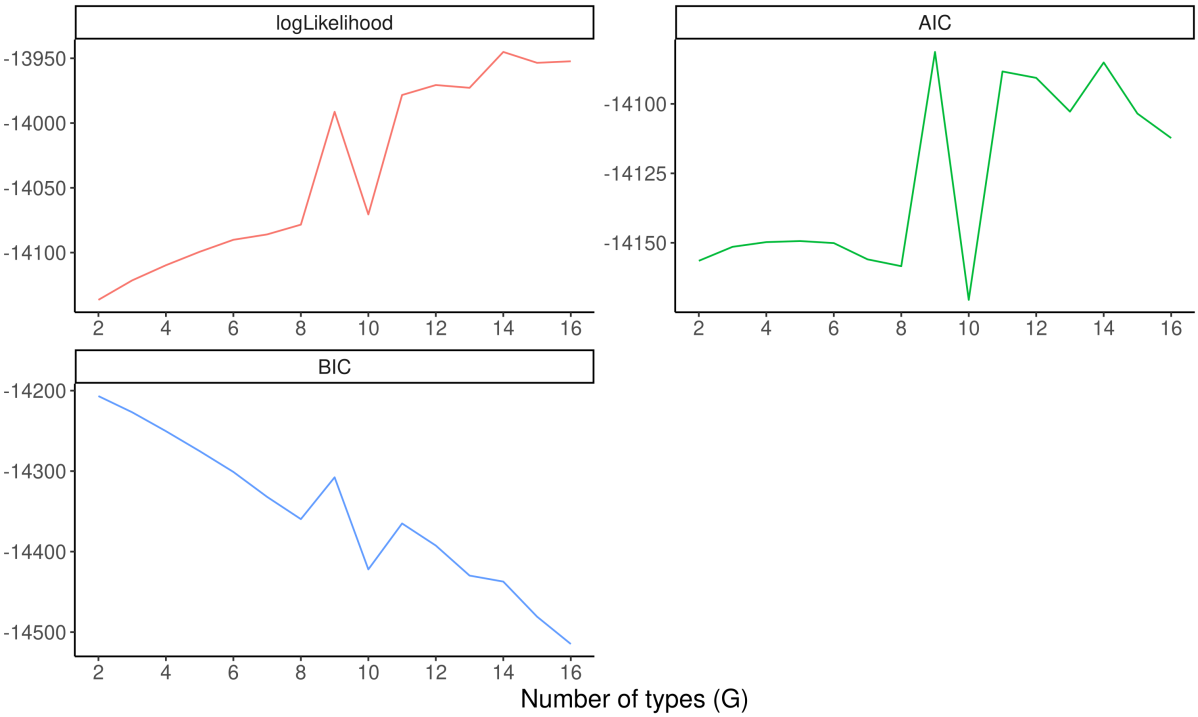
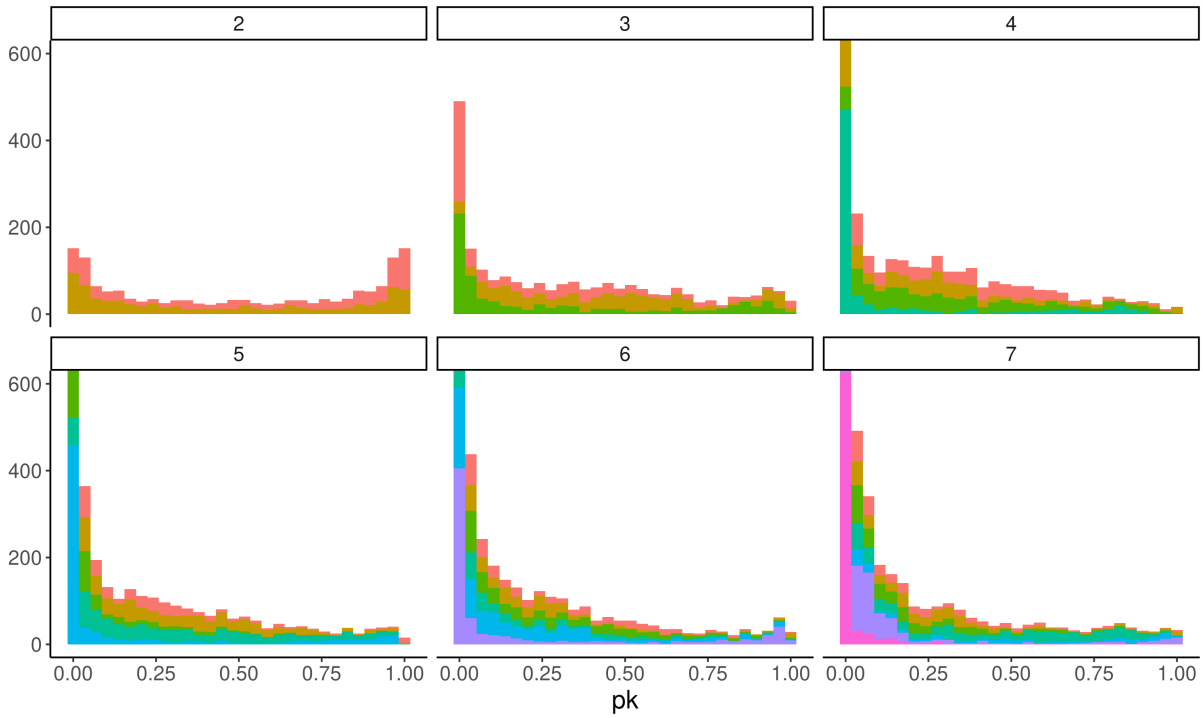


Figure F4: Posterior probabilities: cognitive and non-cognitive measures

(a) Male



(b) Female

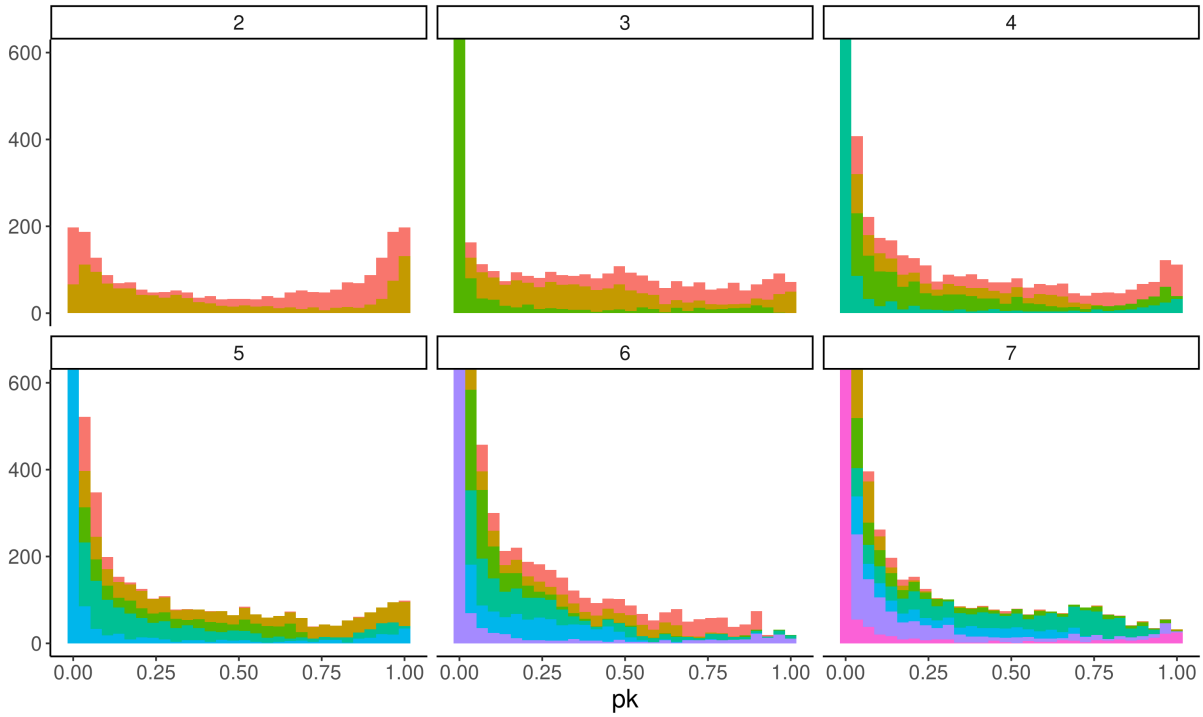
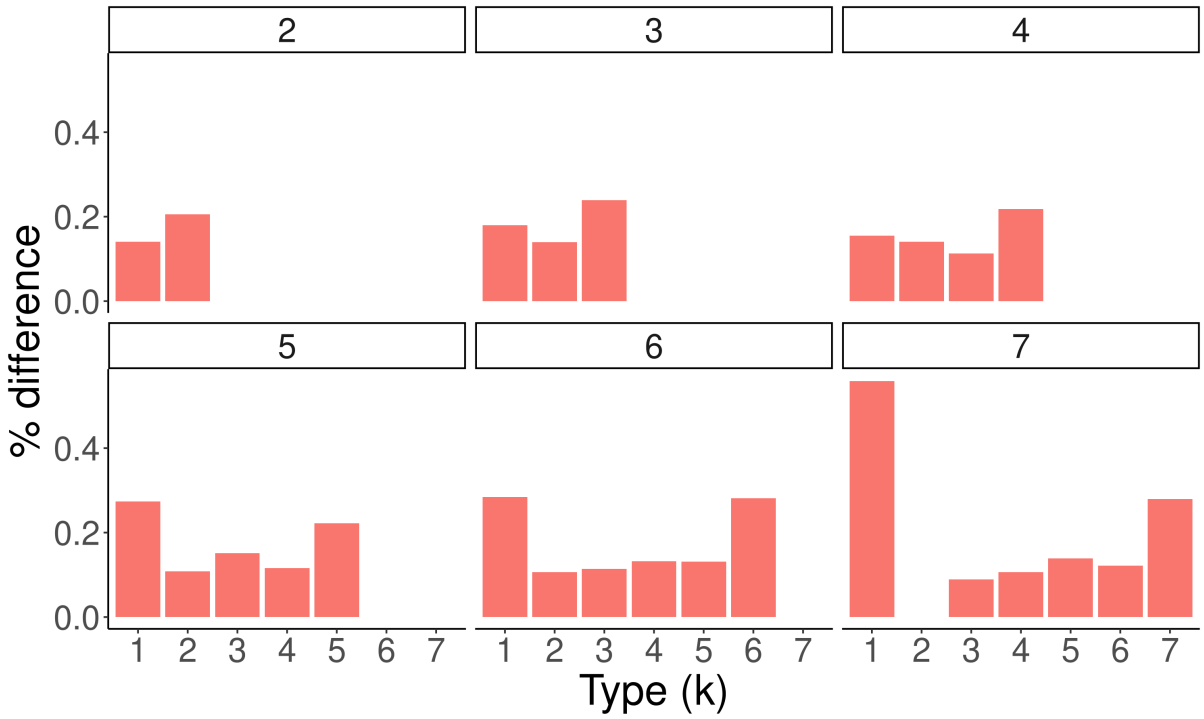


Figure F5: Results across K (single cognitive measure)

(a) Male



(b) Female

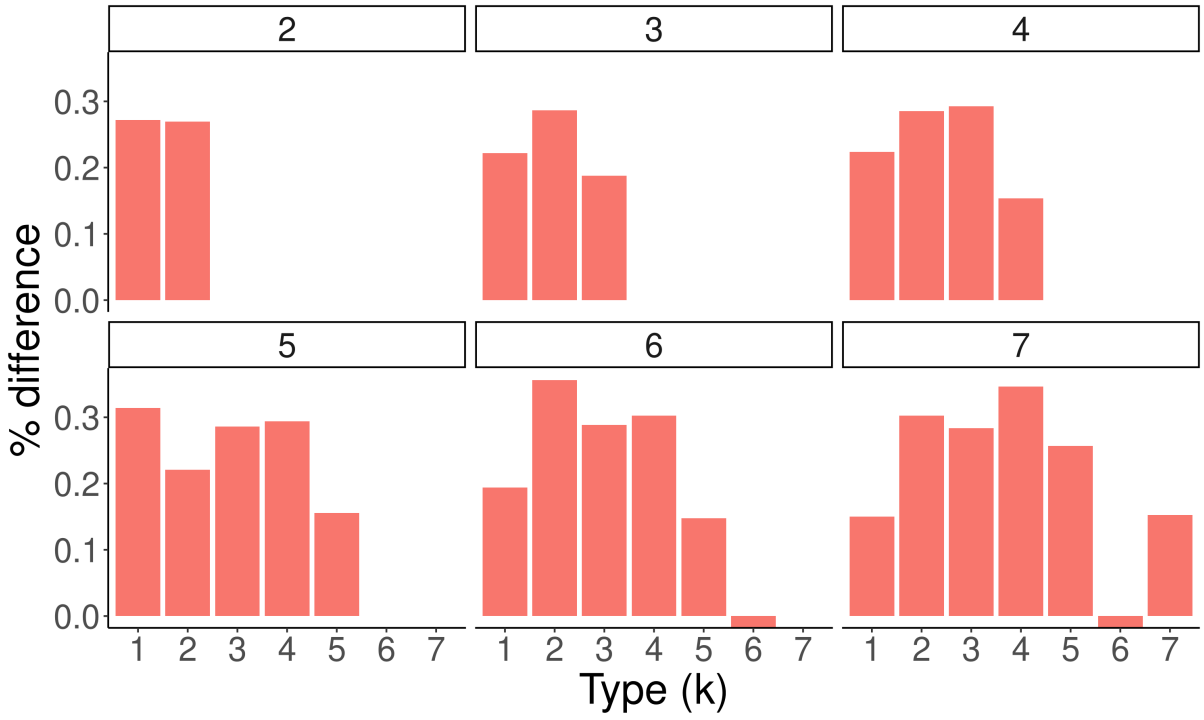
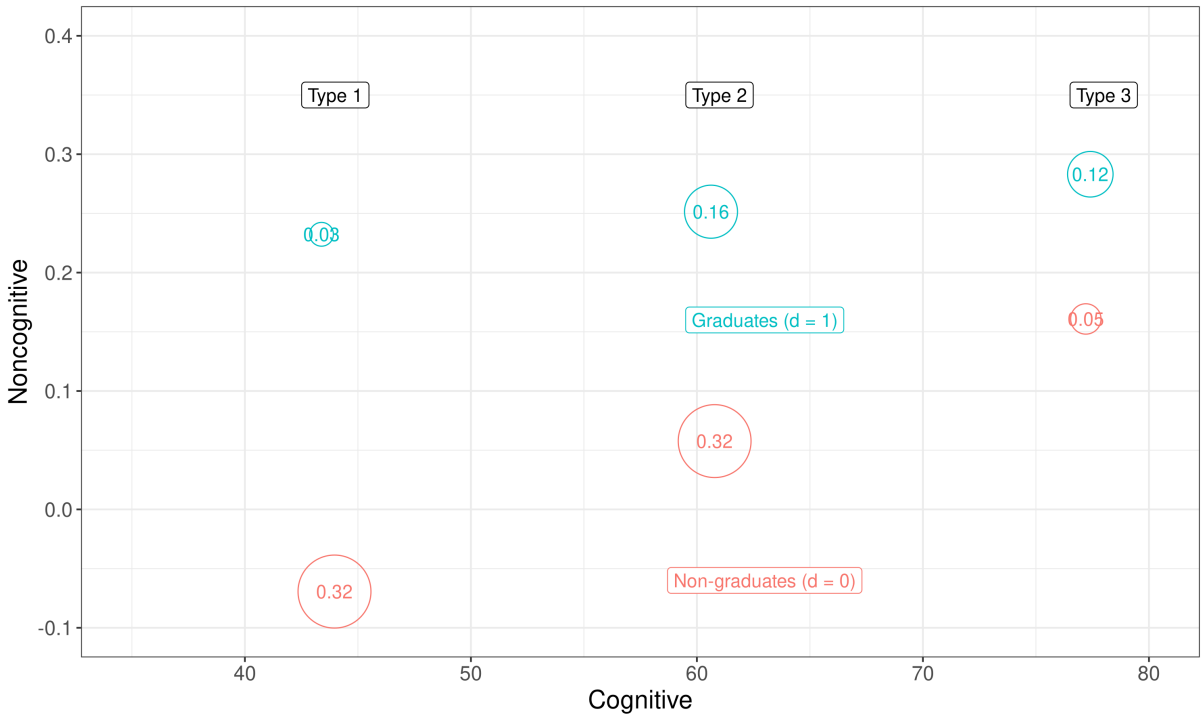
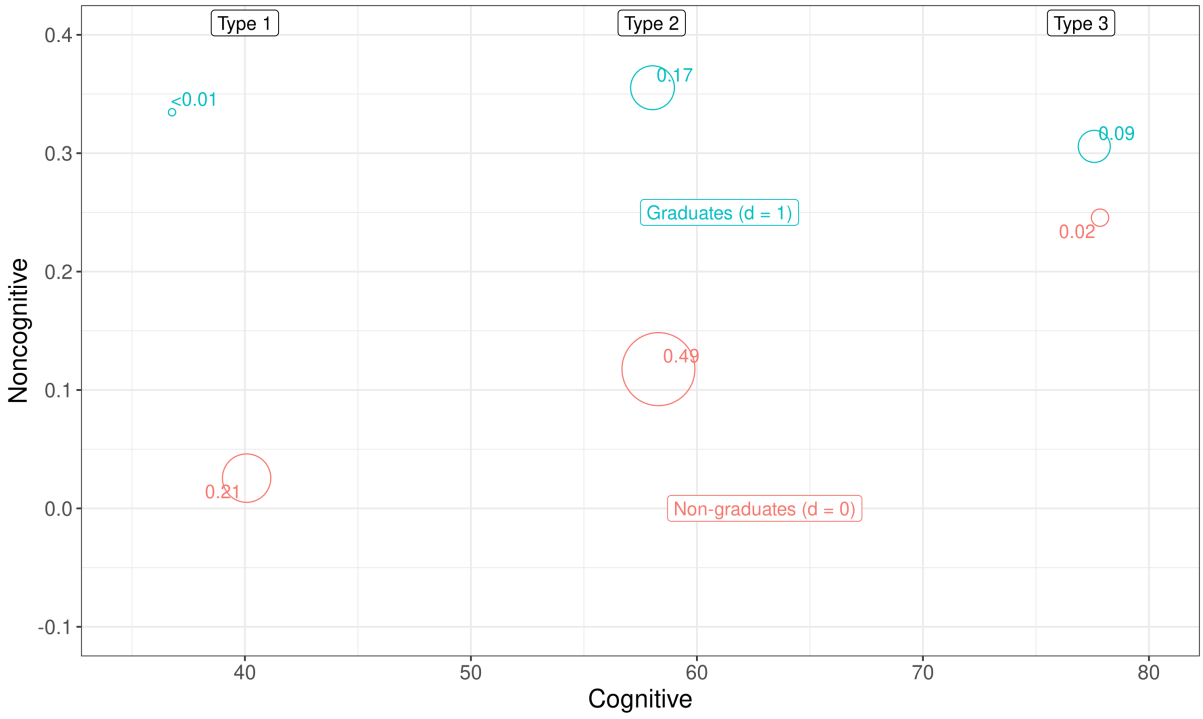


Figure F6: Group sizes and locations in cognitive-noncognitive space ($K = 3$, cognitive measures only)

(a) Male



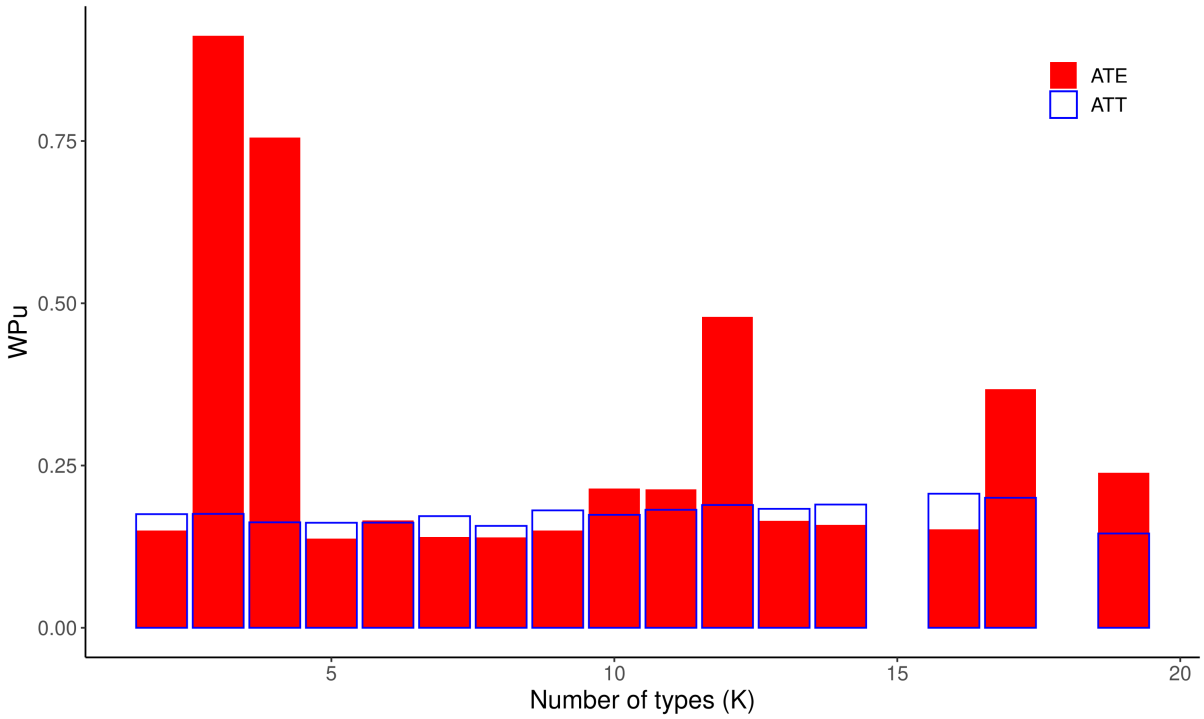
(b) Female



Notes: The above plots display the mean abilities (circle locations) and sizes (circle sizes and labels) for each type, split by gender (panels) and education (colour). Panel (a) contains men, and panel (b) women. Blue circles represent graduates and red non-graduates. The size of each type-education group is labelled, along with each type.

Figure F7: ATEs / ATTs across K (cognitive and noncognitive measures)

(a) Male



(b) Female

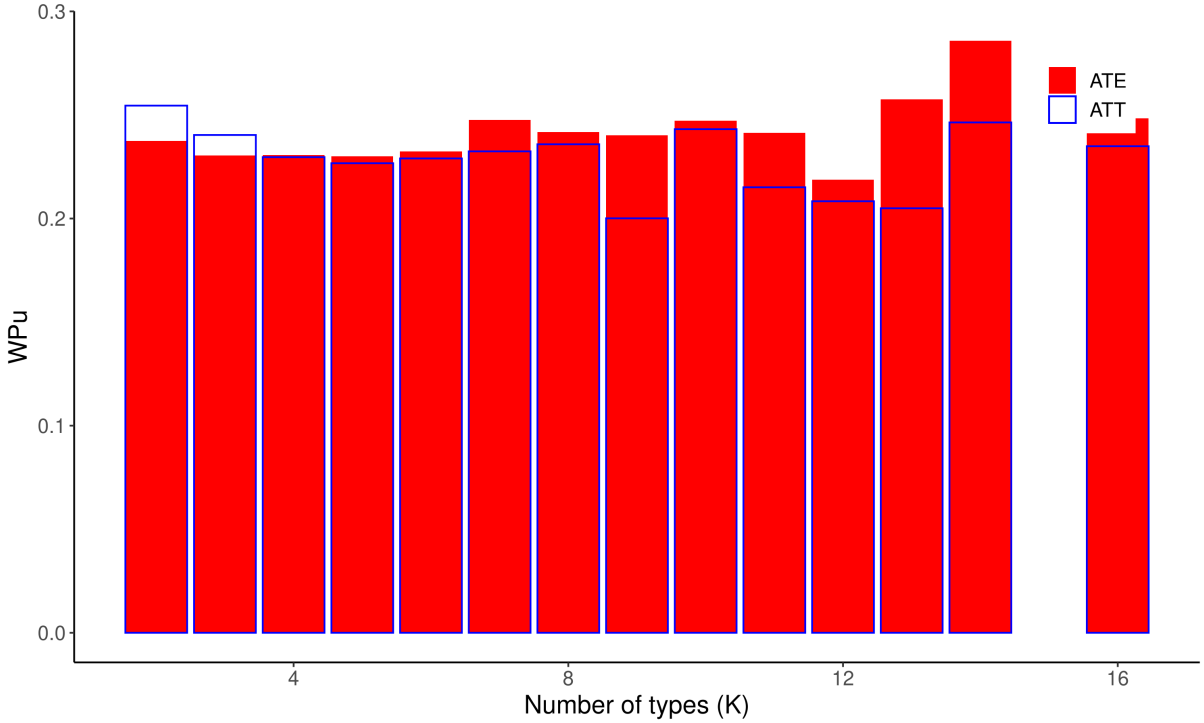


Table F3: $\pi(k,z,d)$ parameter estimates (male, cognitive and non-cognitive)

	$K=5$									
Type$(k)=$	1		2		3		4		5	
$d=$	0	1	0	1	0	1	0	1	0	1
Matters very much	0.004	<0.001	0.045	0.004	0.007	<0.001	0.031	0.028	<0.001	0.022
Matters somewhat	0.072	<0.001	0.119	0.016	0.034	0.004	0.074	0.075	0.015	0.062
Doesn't matter	0.077	<0.001	0.086	0.016	0.015	0.002	0.072	0.046	0.032	0.042

Table F4: $\pi(k,z,d)$ parameter estimates (female, cognitive and non-cognitive)

	$K=5$									
Type$(k)=$	1		2		3		4		5	
$d=$	0	1	0	1	0	1	0	1	0	1
Matters very much	<0.001	<0.001	0.080	0.003	0.019	0.008	0.042	0.027	0.009	0.034
Matters somewhat	0.023	<0.001	0.175	0.007	0.053	0.015	0.065	0.067	0.027	0.046
Doesn't matter	<0.001	<0.001	0.161	0.006	0.041	0.008	0.028	0.026	0.007	0.021

where w_i is log weekly wage, d_i is an indicator for univeristy attendance, M_i^C and M_i^N are cognitive and non-cognitive test scores, X_i contains controls for parental income, location type (city/town/countryside), region, and whether the young person is white, and ε_i is a random error term.

We split the sample by gender and present the results for men in panel (a) and women in panel (b). The first column of table F5 presents the results from the most basic specification, an OLS regression log wages on the degree indicator without any controls. Moving across the columns we add controls to the specification, starting with cognitive in the second column, and non-cognitive (column 3), and then all controls (column 4). Adding controls generally decreases the estimates of the returns to a degree, as one might expect given that wage and university attendance are both positively correlated with prior ability. There is one exception: the coefficient on university attendance when all controls are included for males is larger than with just cognitive and non-cognitive test scores. Finally we use 2SLS with the desire to leave home as an instrument for university attendance, first without (column 5) and then with controls (column 6). The 2SLS estimates are slightly larger than our preferred OLS estimates for men, and much larger for women, suggesting either the strong exclusion restriction required for 2SLS does not hold, or the *compliers* who are induced to attend university by the instrument have unusually high returns (interpreting our 2SLS estimate as a LATE). Recall our main analysis does not require the same exogeneity of the instrument as 2SLS.

Our estimates are broadly in line with previous estimates of the returns to university from the UK during this period. Blundell et al. (2000) estimate a similar equation using OLS with detailed controls on data from a UK cohort born 12 years earlier (in 1958), and using

wages observed later in the life-cycle at age 33. They estimate returns of around 17% for men and 37% for women. We will return to our OLS and 2SLS estimates in section 5 when we use our framework to decompose these estimates using the formulas in section C.

Table F5: OLS and 2SLS estimates of the wage returns to a degree

(a) Male

	Dependent variable: log weekly wage					
	(1)	(2)	(3)	(4)	(5)	(6)
Degree	0.220***	0.181***	0.170***	0.178***	0.207	0.248
	(0.038)	(0.041)	(0.041)	(0.054)	(0.470)	(0.582)
Cognitive		0.003***	0.003**	0.002		0.002
		(0.001)	(0.001)	(0.002)		(0.006)
Non-cognitive			0.057*	0.075*		0.046
			(0.033)	(0.045)		(0.083)
Add. controls				✓		
Instrument					✓	✓
Observations	745	745	745	514	745	745
R^2	0.042	0.052	0.056	0.096	0.042	0.052
Adjusted R^2	0.041	0.050	0.052	0.041	0.041	0.048
Residual se	0.487	0.485	0.484	0.510	0.487	0.486

(b) Female

	Dependent variable: log weekly wage					
	(1)	(2)	(3)	(4)	(5)	(6)
Degree	0.325***	0.291***	0.277***	0.247***	0.530	0.471
	(0.033)	(0.035)	(0.035)	(0.043)	(0.338)	(0.465)
Cognitive		0.003***	0.003***	0.004***		0.001
		(0.001)	(0.001)	(0.001)		(0.004)
Non-cognitive			0.068***	0.034		0.047
			(0.025)	(0.030)		(0.056)
Add. controls				✓		
Instrument					✓	✓
Observations	1,131	1,131	1,131	809	1,131	1,131
R^2	0.078	0.086	0.092	0.150	0.047	0.067
Adjusted R^2	0.078	0.084	0.089	0.118	0.046	0.065
Residual se	0.494	0.492	0.491	0.481	0.502	0.497

Notes: *p<0.1; **p<0.05; ***p<0.01. Specification (1) regresses log-wage on an indicator for a degree and a constant. (2) and (3) include cognitive and noncognitive measures. Then (4) also includes parental income, location type (city/town/countryside), region, and whether the young person is white. Columns (5) and (6) instrument the degree indicator with our instrument.