The returns to higher education by cognitive and non-cognitive abilities

Oliver Cassagneau-Francis*

9th November 2021
(click here for the latest version)

Abstract

Recent work has highlighted the differing returns that individuals from different groups gain from higher education. Combining heterogeneity in returns to higher education with insights on the technology of skill formation from another important recent literature, we estimate the wage returns to a university degree along a fundamental dimension: prior cognitive and non-cognitive human capital. Employing a novel methodology exploiting recent advances in the identification of mixture models, we group individuals by their prior human capital and identify and estimate the wage returns to a university degree by group. Applying our method to data from a UK cohort study, our findings reflect recent evidence that skills and ability are multidimensional. Returns to higher education are large in comparison with the the returns to both cognitive and non-cognitive skills. The returns are also (generally) increasing in ability for both men and women.

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, Kerstin Holzheu, Hugo Lhuillier, Stefan Pollinger, and numerous seminar participants at Sciences Po for helpful comments and discussions. I thank the Centre for Longitudinal Studies at UCL for administering the survey and, along with the UK Data Service, providing access to the data. All errors are mine.

1 Introduction

Over 38% of college-age young people worldwide were enrolled in some sort of tertiary education in 2019, continuing on a steady upward trajectory (from 18% in 1999) which shows no sign of stopping (UNESCO, 2020). As noted by The Economist (2015), "[u]niversity enrolment is growing faster even than demand for that ultimate consumer good, the car." Such demand is perhaps unsurprising, given the raw gap in earnings between university graduates and their less-educated peers remains high across the developed world.[1] Despite this growing demand from across society, higher education remains predominantly a preserve of the rich and privileged (see Blanden and Machin (2004) for the UK, and Chetty, Hendren, Kline, Saez, and Turner (2014) for the USA). High returns to university, coupled with far-from universal attendance, have lead to a widespread belief among economists (and others) that "the component of earnings inequality that is arguably most consequential for the 'other 99 percent' of citizens [is the] dramatic growth in the wage premium associated with higher education and cognitive ability" (Autor, 2014, p. 843).

This powerful statement not only highlights the role of the wage premium for inequality, but also one of the key difficulties in assessing the contributions of higher education to inequality and social mobility: how does one separate the effects of (cognitive) ability from the effects of higher education? With the brightest students attending the best universities, are these universities adding a 40% premium onto the wages these students can command; or would these high ability students have received higher wages even without a degree? We address these questions by proposing a novel framework designed to estimate two key objects: 1) young people's human capital at the time they are deciding whether to attend university; and 2) wages as a function of this human capital and education.

Our method is relatively straightforward to operationalise and is frugal with regards to data: we require fewer measurements of human capital and wages than the current leading approaches, requiring instead an instrument for university (though this need only be exogenous conditional on human capital). We non-parametrically identify the distributions of: human capital before university; graduate and non-graduate wages for *all* young people; the wage returns to a university degree. Armed with this information, we compare the relative contributions of prior ability and higher education to wages and to the dispersion in wages across society. We find that the contribution of higher education to wage inequality is 3 to 4 times the contribution of human capital obtained before university. Our results reflect those of Cawley, Heckman, and Vytlacil (2001) who find that, once education is controlled for, cognitive ability explains little of the variation in wages across individuals even within occupations. Those authors posit that differences in non-cognitive

[1] The mean wage premium across OECD countries was 43.9% in 2017, according to the OECD (2019).

skills may play an important role: however, our results suggest that both cognitive and non-cognitive skills play only a small role in wage inequality.

There is a 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. 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 our attention 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.[2] 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. The framework in this paper incorporates these insights by explicitly including both cognitive and noncognitive skills as determinants of wages, and of the returns to a university degree. Thus we are able to investigate further the direct impact of these skills on wages, while also studying how the returns to education vary across individuals with different levels of cognitive and non-cognitive human capital.

Recent work on the production of human capital during childhood has emphasised the importance of pre-exisiting skills (both cognitive and non-cognitive) in fostering further skills. Cunha, Heckman and coauthors (2006; 2007a; 2008; 2010) in a series of papers identify and estimate the returns to investments in human capital during childhood. They find that skills obtained early in childhood are important for fostering skills later in childhood—a feature they call the *self-productivity* of skills. A related feature is the *complementarity* of skill formation: "skills produced at one stage raise the productivity of investment at subsequent stages" (Cunha et al., 2006, p. 703). These features of skill production during childhood suggest that human capital (skills) at entry to university might affect the effects of a university degree on an individuals skills and abilities—and hence on their later outcomes. A key contribution of our paper is a framework designed to encapsulate this idea by allowing the returns to university to vary with human capital on entry to higher education. We find strong correlation between cognitive and non-cognitive skills, and that graduate wage premia are generally increasing in both types of human capital.

Another recent literature has studied the heterogeneity in returns to a university degree, highlighting that these returns vary considerably across individuals and groups of individuals. Perhaps closest in aim to our paper is recent work by Britton and coauthors

[2] Bowles, Gintis, and Osborne (2001) survey the literature on the determinants of earnings, with a particular focus on noncognitive traits.

(2021a; 2021b). Britton et al. (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 ethinic groups they vary from 7% for White British men, to 40% for Pakistani women. Britton et al. (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. The find weak evidence that returns are positively correlated with the selectivity of the subject or institution. Our paper connects this literature studying the heterogeneity in returns to higher education with the aforementioned literatures on the formation of (multi-component) human capital. We study how the returns to university vary along fundamental dimensions: cognitive and non-cognitive human capital on entry to university.

Another of our key contributions is methodological. We achieve non-parameteric identification using recent advances in the identification of mixture models. Similar techniques have been used to identify a range of models including: firm and worker sorting (Bonhomme, Lamadon, and Manresa, 2019); the wage returns to formal training (Cassagneau-Francis, Gary-Bobo, Pernaudet, and Robin, 2021); and the contributions of workers across different teams (Bonhomme, 2021). Our statistical model, motivated by the literature mentioned above, builds upon the work of Heckman and coauthors (2003; 2006; 2007a; 2010). They typically require strong functional form assumptions to identify their model, assuming an underlying factor model structure (Carneiro et al., 2003). 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. We are able to achieve identification of a similarly flexible model, and we are the first to estimate a non-linear model of this type.[3] Our method is more frugal in its requirements of the data available to the econometrician, requiring half as many measurements to identify models with the same features. We also believe our identification strategy is more transparent, and provide a straightforward estimation strategy via the EM algorithm (an `R` package to implement our method is under development). The cost of this flexibility, frugality, and interpretability is low: we require an instrument for university attendance (though exogeneity is only required conditional on human capital), and assume the distribution of human capital has finite support.

We apply our framework to longitudinal data from an ongoing cohort study in the UK. We first show that a measure of cognitive ability is not sufficient to fully capture variation in human capital across individuals before attending university, despite the 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 human capital on wages, and on the returns

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

to a university degree. The returns to university are also shown to be more important than the returns to 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 human capital 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 3 to 4 times larger than the contribution of ability before university.

The remainder of the paper proceeds as follows. In section 2 we present our model of human capital, wages and higher education. We discuss identification of the additive model in 2.1, and present our novel identification of the non-linear model in 2.3. In 2.4 we discuss some of the usual treatment effect estimators in the context of our framework. We then turn to our application: estimating the wage returns to a university degree as a function of cognitive and non-cognitive human capital. Section 3 discusses the relevant context of higher education in the UK, and presents our data and some initial descriptive results. Section 4 presents the results of estimating our model on this data, first with only cognitive human capital, and then with both cognitive and non-cognitive components. Section 5 concludes.

2 Human capital, education, and wages

Our framework incorporates insights from both the literature on the returns to education, and on the formation of human capital, but we are agnostic about the process by which human capital is formed, both before and during university. Our key period of interest is what we will call the "college years": the period between the end of compulsory education ($t = 0$, age sixteen in our application), and the age at which most people who attend university ($d = 1$ if they attend and graduate, $d = 0$ otherwise) will have graduated and entered the labour market ($t = 1$, at age twenty six in our application). Young people possess some human capital at the time they make the decision to attend university ($t = 0$). We will call this their "human capital endowment" and denote it by θ.

We do not observe θ. This is problematic for our aim of estimating the wage returns to university. First, because human capital 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. Second, it is likely that the returns to university also vary with θ as university is an investment in human capital and recent work has highlighted the role of existing human capital in fostering additional human capital. To deal with these issues, we will use noisy "measurements" of θ to proxy for the true values. Using the information on θ contained in these measurements and in wages, we will be able to estimate the distribution of θ, and study how the returns to university vary across this distribution.

Wages, W. We observe wages for individuals with education d in $t=1$, W_d. Log wages are a function of an individual's human capital endowment, θ, education, d, and some error term, ϵ:

$$\ln W_d \equiv w_d = \mu_d(\theta, \epsilon_d)$$

The error term contains other (unobserved) factors that affect wages beyond education and human capital. Human capital can be multidimensional, consisting of, for example, cognitive and non-cognitive abilities. Then $\theta = (\theta^C, \theta^N)$.

Each individual therefore has two *potential* wages, of which we only observe one. This is the classic problem in the potential outcomes framework, or in estimating treatment effects. We are interested in estimating (the distribution of) $w_1 - w_0$, the effect of attending university on wages.

Economic motivation for the model. We can motivate our statistical model by drawing from the literature on human capital formation. Consider a simple version of the 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 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.

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

Human capital can have multiple components, for example capturing cognitive and non-cognitive abilities, both of which have been shown to be important for fostering further human capital and (socio-economic) outcomes. In the multi-component case θ_t is a vector: $\theta_t = (\theta_t^1, \theta_t^2, ..., \theta_t^J)$.

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 discrete.

$$\theta_1 = f_\theta(\theta, d)$$

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

$$w_{d1} = w(\theta_1, \epsilon) = w(f_\theta(\theta, d), \epsilon) = \mu_d(\theta, \epsilon_d).$$

Measurements, M_ℓ. We do not directly observe human capital endowments, but have access to noisy measurements of (the components of) θ. We observe L measurements of human capital in $t=0$. These might be test scores measuring cognitive ability, or behaviour or personality scales which also depend on noncognitive abilities. Each measurement is a function of (some component(s) of) human capital. Without loss of generality, we can set one measurement to only depend on one factor. We will do this for the measure of cognitive ability (reading and mathematics tests in our application), and call this the "cognitive" component, θ^C. We will call the other the non-cognitive component. All other measurements depend on both components.

$$M_C = m_C(\theta^C, \epsilon_C)$$
$$M_\ell = m_\ell(\theta^C, \theta^N, \epsilon_\ell),\ \forall \ell \neq C.$$

The error terms ϵ contain other (unobserved) factors that might affect the measurements beyond human capital.

2.1 Additive wages and measurements

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

$$w_d = \beta_d + \beta_d^C \theta^C + \beta_d^N \theta^N + \varepsilon_d \tag{2}$$
$$M_\ell = \gamma_\ell^0 + \gamma_\ell^C \theta^C + \gamma_\ell^N \theta^N + \varepsilon_\ell. \tag{3}$$

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

Assumption 1 (Independent errors)**.** *The error terms, ε_d and ε_ℓ are independent of θ and each other, and are mean zero.*

Under assumption 1 we obtain the *classical measurement error model*, and OLS estimates using M to proxy θ as in the following equation,

$$w_d = \delta_d^0 + \boldsymbol{M'} \delta_d + \eta_d \tag{4}$$

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 (2) and (3) to match equation (4) and equating the error terms. The second equality follows from assumption 1. Therefore, we cannot recover θ nor any of the parameters in equations (2) and (3) via OLS. However, models with measurment error are well studied in economics and statistics. When w and M are *not* jointly normal, Reiersol (1950) shows that the parameters in equations (2) and (3) 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).[4] Bonhomme and Robin (2009, 2010) generalise these results to allow for the non-parametric identification and estimation of such factor models.

The linear model allows us to study the heterogeneity in returns to education, and to compare the contribution of education to wage dispersion with the contribution of prior human capital. We can also study the correlation between cognitive and noncognitive abilities. However, the linearity assumption shuts down any complementarities between components of human capital both on wages directly, and on the returns to education. CH's approach also requires homoscedasticity: errors cannot depend on human capital. We are able to relax this assumption for both the wage and measurement equations. The novel framework introduced in this paper allows the nonparametric identification of the distributions of wages and measurements—i.e. without imposing any parametric or functional form assumptions.

2.2 Allowing wages and measurements to be nonlinear

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 human capital for identification than CH's approach, and we are able to identify and estimate a fully non-linear model.[5] This frugality and flexibility comes at a low cost: compared with CH we additionally require: 1) an instrument for university attendance, which we denote z, and which need only be exogenous of wages and measurements conditional on human capital; 2) that the distribution of human capital has finite support.

Our assumption that the distribution of human capital has finite support, implies that (each component of) θ takes one of K values:

$$\theta \in \{\theta_1, ..., \theta_k, ..., \theta_K\}.$$

When θ has only one component, there is a natural ordering on θ: $\theta_1 < \theta_2 < ... < \theta_K$. However, with a multi-component θ there is no such ordering. For example, in our application we assume cognitive and non-cognitive components of human capital, $\theta_k = (\theta_k^C, \theta_k^N)$, and we place no restrictions on the relative ordering of the components. Therefore it is possible to have $\theta_1^C < \theta_2^C$, but $\theta_1^N > \theta_2^N$. We will call the index k on θ_k an individual's

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

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

"type" and we label types so that k is increasing in expected non-graduate wages, i.e.

$$\mu_0(\theta_1) < \mu_0(\theta_2) < \ldots < \mu_0(\theta_K). \tag{5}$$

We will assume during our discussion of identification that the econometrician knows the *true* number of points of support, K. However, we will also discuss how this can be estimated when we operationalise the method.[6] Our aim is to identify the points of support for the distribution of human capital, θ, the distribution of human capital, $\pi(\theta)$, the distribution of wages conditional on human capital and education, and the distributions of the measurements, conditional on human capital. We introduce (the notation for) these objects now.

Wages at $t = 1$ after education d, W_d. Wages of an individual with human capital θ and education d are distributed as

$$\ln W_d = w_d \sim f_w(\cdot|\theta, d).$$

Measurements at $t = 0$, M_ℓ. The distribution of each measurement depends only on human capital.

$$M_\ell \sim f_\ell(\cdot|\theta),$$

Individual measurements and wages are not directly informative of human capital due to noise in these measures / outcomes. However, under our maintained assumptions, the mean wages and measurements across individuals with the same θ *are* informative. Therefore we can use these means to the set the support of θ in an interpretable metric, and also to compare across *types*, i.e. groups of individuals with the same level of human capital, θ.

A key requirement for identification is an instrument for education. Our exclusion restriction on the instrument is less restrictive than for a classic instrumental variable, as z need only be independent of wages and measurements conditional on human capital (and education for wages).

Instrument for education. We denote the joint probability of observing education d, human capital θ, and instrument, z by $\pi(\theta, z, d)$.[7]

[6] We will also discuss to what extent our method can be used to estimate discrete approximations of a continuous distributions.

[7] An economic motivation for our setup, and in particular the role of the instrument, is the extended Roy model (Heckman and Vytlacil (2005); Carneiro, Heckman, and Vytlacil (2010, 2011)):

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

2.3 Non-parametric identification of the model

Likelihood of an individual's observations. Under the model detailed in the previous subsection, the likelihood associated with an individual i's observations writes

$$\ell(\boldsymbol{M}_i, w_i, d_i, z_i) = \sum_{\theta} \pi(\theta, z_i, d_i) f_M(\boldsymbol{M}_i|\boldsymbol{k}) f_w(w_i|\boldsymbol{k}, d) \tag{6}$$

with $\boldsymbol{M} = (M_1, ..., M_L)$, $f_M(\boldsymbol{M}|\theta) = \prod_\ell f_\ell(M_\ell|\theta)$.

We use results from recent work on mixture models to show the elements on the right-hand side of equation (6) are identified under the following assumptions (see Bonhomme, Jochmans, and Robin (2017) for details).

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

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

Assumption 3 (Linear independence)**.** $\left[f_M(\boldsymbol{M}|\theta_1) \quad \cdots \quad f_M(\boldsymbol{M}|\theta_k) \quad \cdots \quad f_M(\boldsymbol{M}|\theta_K)\right]$ *and, for all* d*,* $\left[f_w(w_i|\theta_1, d) \quad \cdots \quad f_w(w_i|\theta_k, d) \quad \cdots \quad f_w(w_i|\theta_K, d)\right]$ *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 correpsonding to other points of support. This is analogous to the rank condition in ordinary least squares.

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

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

Assumption 5 (Discrete wages and measurements)**.** *The distributions of wages and measurements have discrete support.*

Assumption 5 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

where θ is human capital (different social backgrounds, as measured/influenced by control variables such as parental income, cognitive and noncognitive ability etc, produce different levels of human capital). 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. $w(\theta, 0), w(\theta, 1)$ are treatment-specific outcome variables (random given θ and independent of z). $c(\theta, z)$ is cost of attending university (pecuniary and non-pecuniary, random given θ, z).

to adapt the proof.

Theorem 6 (Identification)**.** *Under assumptions 2-5 plus the conditional exclusion restriction on the instrument,* $\pi(\theta, z, d)$*,* $f_M(\boldsymbol{M}|\theta) = \prod_\ell f_\ell(M_\ell|\theta)$*, and* $f_w(w|\theta, d)$ *are non-parameterically identified.*

The proof of theorem 6 is in appendix A. Assumptions 2-5 ensure that the matrix containing the (discrete) joint probability distribution of the observed variables—$P(\boldsymbol{M}, w, z, d)$—is rank K. The proof then uses manipulations of the singular value decomposition of this matrix to show that the elements on the right hand side of equation (6) are uniquely determined, (see similar arguments in Kasahara and Shimotsu, 2009; Hu and Shum, 2012; Bonhomme, Jochmans, and Robin, 2016a,b; Bonhomme et al., 2019; Cassagneau-Francis et al., 2021).

2.4 Treatment effects in our framework

We can calculate some of the usual objects that analysts hope to estimate when considering the returns to education—average treatment effect (ATE), average treatment on the treated (ATT)—and demonstrate some of the biases that more standard estimation approaches suffer from. We are also able to calculate heterogeneous treatment effects, conditional on the level of human capital a person possesses on entry to university. 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 have reason to believe that wages are correlated with education—which is likely as higher human capital results in both higher wages *and* a increased probability of attending university—then

$$\mathbb{E}[w_1 - w_0] \neq \mathbb{E}[w_1 \,|\, d = 1] - \mathbb{E}[w_0 \,|\, d = 0]. \tag{7}$$

This is the classic problem of *selection into treatment.*

Fortunately, our maintained assumptions provide a solution. Potential outcomes are independent of education and z conditional on human capital, i.e. $w_1, w_0 \perp\!\!\!\perp z, d \,|\, \theta$. The observed outcome, $w = dw_1 + (1 - d)w_0$. We can then define the following treatment effects.

Average treatment effect by human capital, $ATE(\theta)$**.** This is the expected wage gain from a university degree for a young person with human capital level θ.

$$ATE(\theta) \equiv \mathbb{E}[w_1 - w_0 \,|\, \theta] = \mathbb{E}[w_1 \,|\, \theta] - \mathbb{E}[w_0 \,|\, \theta]$$

Average treatment effect. We can aggregate over types to obtain the ATE in (7).

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

where $\pi(\theta) = \sum_{z,d} \pi(\theta, z, d)$, the proportion of individuals with human capital level θ.

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

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

where $\pi(\theta \,|\, d = 1) = \frac{\sum_z \pi(\theta,z,1)}{\sum_{\theta,z} \pi(\theta,z,1)}$, the proportion of individuals with human capital level θ 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}.

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

where

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

The OLS bias disappears only if (i) $\pi(\theta \,|\, d = 1) = \pi(\theta \,|\, d = 0)$ for all values of θ; or (ii) $\mathbb{E}[w_0 \,|\, \theta] = \mathbb{E}[w_0 \,|\, \theta']$ for all $\theta \neq \theta'$. The first equality is unlikely to hold in our application as those with higher human capital are more likely to attend university, and hence the proportion of those with high human capital 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 human capital 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, 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

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

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

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

where

$$LATE = \sum_{\theta} \frac{\pi(\theta, d=1|z=1) - \pi(\theta, d=1|z=0)}{\sum_{\theta}\left[\pi(\theta, d=1|z=1) - \pi(\theta, d=1|z=0)\right]} ATE(\theta), \tag{8}$$

and

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

with

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

Therefore the $LATE$ estimator of Imbens and Angrist (1994) is a weighted average of θ-specific ATEs in our framework, with weights corresponding to the proportion of *compliers*[8] with that level of human capital. Note the similarity between our decomposition of the $LATE$ in equation (8) and Heckman and Vytlacil (1999, 2005, 2007)'s marginal treatment effect (MTE):[9]

$$LATE = \frac{\int_{\varphi(0)}^{\varphi(1)} \Delta^{MTE}(\nu)\,\mathrm{d}\nu}{\varphi(1) - \varphi(0)} \tag{9}$$

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 (9) 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 (unobserved) human capital. Our framework is more flexible in one sense as it allows correlation between outcomes (w_d) and unobserved costs through θ.

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

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

3 Application: higher education in the UK

We now turn to our application, in which we estimate the returns to a university degree in the UK. The data comes from the British Cohort Study, an ongoing longitudinal study of all children born in the UK in a particular week in April 1970. This study, one of a number of similar longitudinal cohort studies in the UK, is particularly suited to demonstrate our method as it contains detailed information on the cognitive and non-cognitive skills from before university, and wages and occupation data from after they (would) have graduated from university. We first briefly describe the context of higher education in the UK, then discuss the data and the specific variables used to estimate the model, and the results are in section 4.

3.1 Higher education in the UK

The higher education system in the UK is particularly suited to estimating the human capital production model in this paper for a number of reasons. First, the institutions and the degrees they offer are relatively homogenous with all degree-granting institutions being privately-run though receiving government funding. The standard degree offered by a UK university is a three-year bachelors degree specialising in a single subject. The student would then be awarded a Bachelor of Arts (BA, in arts or humanities subjects) or a Bachelor of Sciences (BSc) in that subject upon graduation. Most universities offer students a wide range of subjects, and have large student bodies—the largest have nearly 30,000 students. 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). Our choice of instrument is quite particular to the UK, as 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).

3.2 Data

The main source of data is the British Cohort Study (BCS) 1970, an ongoing longitudinal cohort study of every person born in the United Kingdom in a week in April 1970. There were approximately 16,000 initial cohort members, who have been followed and contacted roughly every five years since their birth, with eleven "sweeps" to date. The latest sweep is currently underway in 2021 (the cohort members are aged 51). In each sweep the contactable cohort members (and / or their family) were interviewed about their current circumstances and daily life, with more specific focuses at different stages of their lives. Relevant for our analysis 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.

Therefore our analysis focuses on the fourth wave, which took place in 1986 (when the cohort members were aged 16), and on the fifth wave which took place a decade later in 1996 (aged 26). These are key because they provide information from just before the decision to attend university, which is generally made at age seventeen in the UK, and from "after university", when most young people who would attend university have completed their degrees and entered the labour market. We 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. We leave the discussion and investigation of this gender pay gap to other work.

Table 1 contains summary statistics for the variables used to estimate our model, for the whole sample and then by education and gender. Wages are increasing in education for both men and women. The mean wage for graduate women is below that of non-graduate men, despite women having similar cognitive test scores and higher noncognitive measures, confirming our decision to estimate the model separately for men and women. Cognitive and noncognitive measures are positively correlated with education for both men and women. Our cognitive measure is the mean standardised score (out of 100) across reading and mathematics tests taken by the cohort members as part of the study at sixteen. Our non-cognitive measure is a standardised score across three measures of "personality": self esteem, locus of control, and the general health questionnaire (GHQ).[10] Our instrument, a measure of how strongly an individual wishes to leave home, is positively correlated with holding a degree at age twenty five.

Table 2 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 sixteen. The results are reassuring, as the majority of the coefficients on the instrument are not significant, even at the 10% level, except self-assessed health at age sixteen. Table C9 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 human capital using our measurements.[11]

[10]The GHQ is a series of questions designed to predict susceptability to mental health issues. Conti and Heckman (2010) use similar measures from the same dataset to capture non-cognitive abilities and their effects on later health outcomes.

[11]As a further check on the robustness of our method to different instruments we plan to estimate our model using a more standard instrument: the distance to the nearest university at age sixteen. However,

Table 1: Summary statistics for the full sample and by sex and education

	All	Male		Female	
		$D=0$	$D=1$	$D=0$	$D=1$
Weekly wage (age 25, GBP, W)					
Mean	239	263	334	197	241
Std dev.	408	481	474	388	212
Degree	0.29	0	1	0	1
Male	0.40	1	1	0	0
Ability measures (M)					
Cognitive	57.0	54.3	65.3	53.6	64.4
Noncognitive	0.13	0.01	0.26	0.10	0.34
Leaving home matters... (z)					
...very much	0.19	0.13	0.17	0.20	0.27
...somewhat	0.48	0.46	0.50	0.47	0.50
...doesn't matter	0.33	0.41	0.34	0.32	0.23
N	1876	509	236	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.

The balancing exercise suggests the wish to leave home is uncorrelated with other characteristics that might determine wages conditional on human capital, and appears to be a valid instrument for our purposes.

In table 3, 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

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

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 on males in panel (a) and females in panel (b). Column (1) of table 3 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 (2), and non-cognitive (3), and then all controls in (4). Adding controls generally decreases the estimates of the

we are yet to obtain access to location data at the finer level required to calculate this distance.

Table 2: 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 we use ordered logit 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.

returns to a degree, as one might expect given that wage and university attendance are both positively correlated with human capital. 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 wish to leave home as an instrument for university attendance, first without (5) and then with controls (6). The 2SLS estimates are slightly larger than our preferred OLS estimates for men, and much larger for women, suggesting either the strong exlclusion 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).

Our estimates are broadly in line with previous estimates of the returns to university from this period. Blundell, Dearden, Goodman, and Reed (2000) estimate a similar equation using OLS with detailed controls on data from a 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 4 when we use our framework to decompose these estimates using the formulas in section 2.4.

3.3 Estimation details

Although our non-parametric identification strategy detailed in section 2.3 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.

We assume that the conditional measurements normally distributed conditional on human capital, and that wages are log-normally distributed conditional on human capital and education. Therefore, measurement M_j has probability density function (PDF)

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

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

Similarly, log-wages, w, are distributed as

$$f(w|\theta, d) = \phi\left(\frac{w - \mu(\theta, d)}{\sigma(\theta, d)}\right)$$

We can now use Dempster, Laird, and Rubin (1977)'s expectation-maximisation (EM) algorithm to estimate the parameters of the model via maximum likelihood. The computuational burden can be further reduced by applying Arcidiacono and Jones (2003)'s

Table 3: 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				Yes		Yes
Instrument					Yes	Yes
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				Yes		Yes
Instrument					Yes	Yes
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, without and including controls.

sequential-EM algorithm which avoids having to estimate many parameters in one step.

The maximum likelihood estimator of the parameters, Ω, satisfies

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

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

The sum inside the logarithm prohibits sequential estimation of the parameters in Ω. Arcidiacono and Jones (2003) show the same $\hat{\Omega}$ satisfies

$$\hat{\Omega} \equiv \sum_{i=1}^{N} \sum_{\theta=1}^{K} p_i(\theta|\Omega) \ln \ell(\Omega; \boldsymbol{M}_i, w_i, z_i, d_i, \theta) \tag{10}$$

where

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

and

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

Crucially, the right-hand side of (10) *lends itself to sequential estimation.*

We use the sequential EM algorithm to maximise (10), using `R`'s *kmeans* function on $\boldsymbol{m}$ and w to select starting values.[12] Details of the algorithm are in appendix B.[13] The variables we use as w, $\boldsymbol{M}$, z, and d are detailed in table 1. 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. These criteria are discussed in the next section.

4 Results

This section presents the results of estimating our model on the data as described in the previous sections. We first discuss how to choose the number of points of support for the distribution of human capital, K. We then present results by human capital level using only cognitive measurements, which is not our preferred specification but as many datasets do not contain noncognitive so it is informative to see how the model performs with only this information. We then turn to our preferred specification which includes measures for

[12]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.

[13]The algorithm is relatively fast to converge, taking under one minute on a laptop with a quad-core Intel Core i7-6560U CPU (2.20GHz) processor and 16GB of memory, running Fedora (Linux).

both cognitive and noncognitive human capital, and finally we compare aggregate results across specifications and to estimates obtained using standard approaches.

Definition 4.1 (Types)**.** *We refer to individuals with the same level of human capital as a "type", and we will label these types according to the index on θ. Therefore type k individuals are those with human capital θ_k. A scalar type, k, can capture multidimensional human capital, $\theta_k = (\theta_k^C, \theta_k^N)$.*

4.1 Choosing K

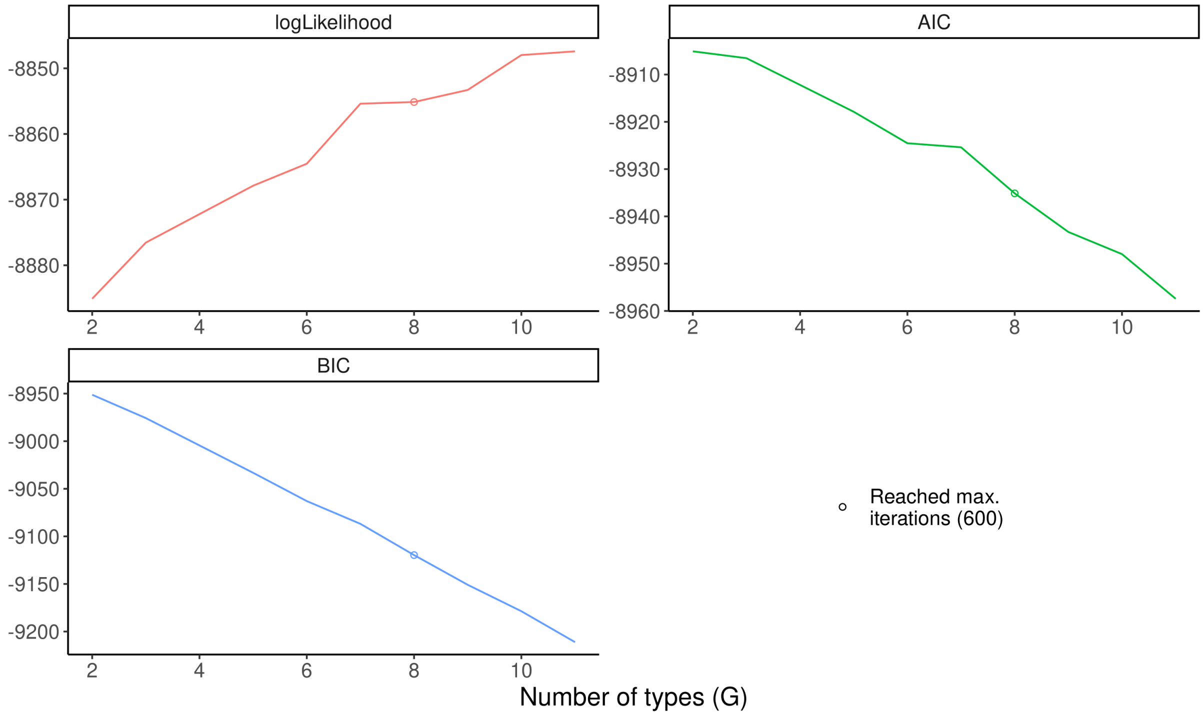
We use a range of criteria to select the number of points of support for the distribution of human capital or "types". The *likelihood criteria*, displayed in figure 1 for the model with only a cognitive measure estimated on men, are the log-likelihood, and penalised log-likelihood for different values of K. We are looking for either elbows where the slope of the plotted line decreases (all) or maxima (AIC, BIC). The plots in figure 1 suggest picking a value of K less than 7 (though BIC is uninformative). We also look at the results across different K to see if there are any clear patterns, or whether certain values of K appear to produce anomalous results. Generally we select the lowest value of K which appears to capture the key patterns in the returns to university, as lower K are preferred by the criteria in figure 1. We can also study the "entropy" of the assignment to groups: the uncertainty or "fuzziness" of the assignment, which we can measure by the distribution of the posterior probabilities, $p_i(k)$. Stronger assignments will have modes of the posterior probability distribution at 0 and 1. Figure 2 displays the distribution of posterior probabilities for the same model as figure 1 though this time 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 in the appendix.

4.2 Cognitive measure of human capital

We first present our estimates using a single cognitive measure. Although not our preferred specification, there are many datasets containing only measures of cognitive ability and these are generally much more widespread than those containing noncognitive measures (or both). Therefore understanding how close to our "ideal" specification we can get using only cognitive measures is an important empirical question.

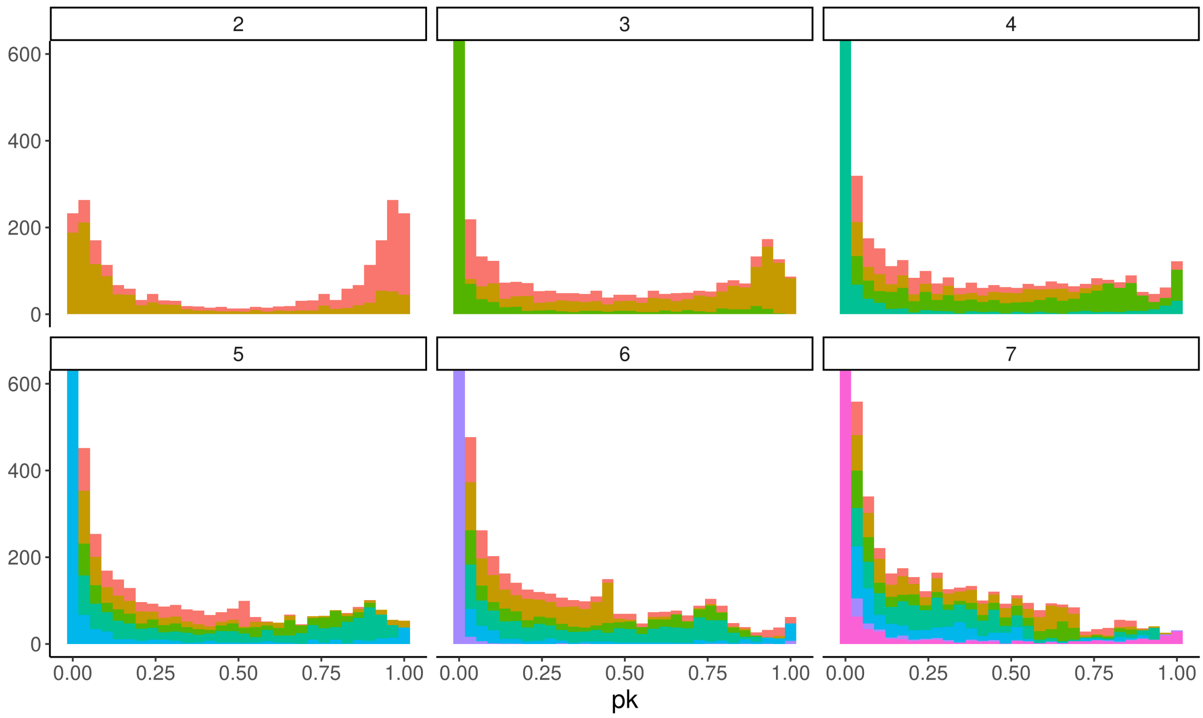
Table 4 displays the results for three levels of human capital (types). The results for other numbers of types, which are broadly similar, are in the appendix (figure D10). The types are labelled so that k is increasing in the mean wages of those without a degree, $\mu(\theta_k, d)$, although with a single measure this is the same ordering as if we had used the mean of the cognitive measurement, $\alpha_C(\theta_k, d)$. By estimating the model separately for men and women, we may estimate a different set of types for men and women (corresponding to

Figure 1: 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 when the algorithm had not converged within 400 iterations.

Figure 2: Distribution of posterior probabilities (cognitive measure, females)



Notes:

Table 4: Results by level of human capital with a single cognitive measure ($K = 3$)

	Male						Female					
$k =$	1		2		3		1		2		3	
Degree	0	1	0	1	0	1	0	1	0	1	0	1
Return to a degree	0.179		0.140		0.239		0.222		0.286		0.188	
Wage (age 25, GBP)												
Mean	205	246	221	254	230	292	147	184	158	210	188	227
Ability measures, $\mathbb{E}[M_j \vert \theta_k, d]$												
Cognitive	44.0	43.4	60.8	60.6	77.2	77.4	40.1	36.8	58.3	58.0	77.8	77.6
Non-cognitive	-0.07	0.23	0.06	0.25	0.16	0.28	0.03	0.34	0.12	0.35	0.25	0.31
$\pi(k)$	0.32	0.03	0.32	0.17	0.05	0.12	0.21	<0.01	0.49	0.17	0.02	0.09

different distributions of human capital). However, we can compare types within and across genders using the type-conditional means, $\alpha_\ell(\theta_k)$ and $\mu(\theta_k, d)$.[14]

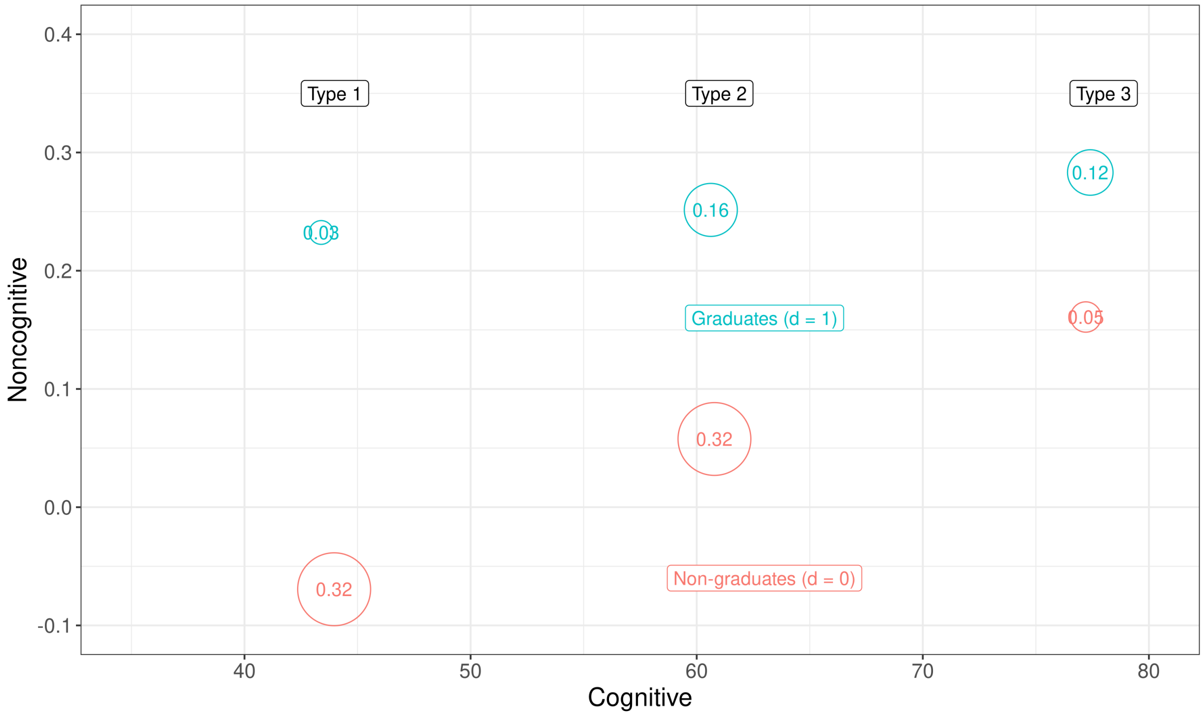
Although there is a significant gender wage gap, the estimated types are similar in terms of cognitive ability across males and females. For example type 1 men have a mean cognitive score of approximately 45, while type 1 women have a mean cognitive score of 40. Despite having lower wages and similar cognitive scores, the mean noncognitive scores of the female types are higher than those for the equivalent male type. This highlights the importance of studying men and women separately as they likely face different prices for their human capital on the labour market. Table 4 also splits the type-conditional means by education, d. For the cognitive measures which is used to estimate types, the mean functions appear to be independent of education. However, without including a measure of noncognitive skills in the specification, noncognitive ability is correlated with education within types.

Returns at each level of human capital are generally higher for women then men, except for those with the highest cognitive ability. The pattern of returns across prior (cognitive) human capital also differs across genders. The pattern for men is U-shaped, with those in the middle of the human capital distribution experiencing lower returns than those with high or low cognitive ability. For women returns are hump-shaped with respect to cognitive human capital, with middle types receiving the highest returns. These shapes are more apparent in figure D10 in the appendix. These non-linearities in returns with respect to cognitive ability emphasises the importance of a framework such as ours which does not impose additivity. However, given the differences between men and women in terms of noncognitive ability, and the correlation within types between noncognitive 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.

This difference is made clear in figure 3 which displays each type in the space of cognitive and non-cognitive skills. For both men (panel (a)) and women (b) there are large differences within types in terms of non-cognitive skills between graduates (blue) and non-graduates (red). We can consider our method as a type of *matching estimator*, comparing the outcomes of individuals similar in human capital at age 16 some of whom attend university, while others do not. Figure 3 shows that when we only use a cognitive measure, we are only successful in matching along the cognitive dimension. Therefore, despite correlation between cognitive and non-cognitive skills it appears important to include a measure of non-cognitive skills in the model. We do so in the following section.

Figure 3: Types in cognitive-noncognitive space ($K = 3$, cognitive measure)

(a) Male



(b) Female

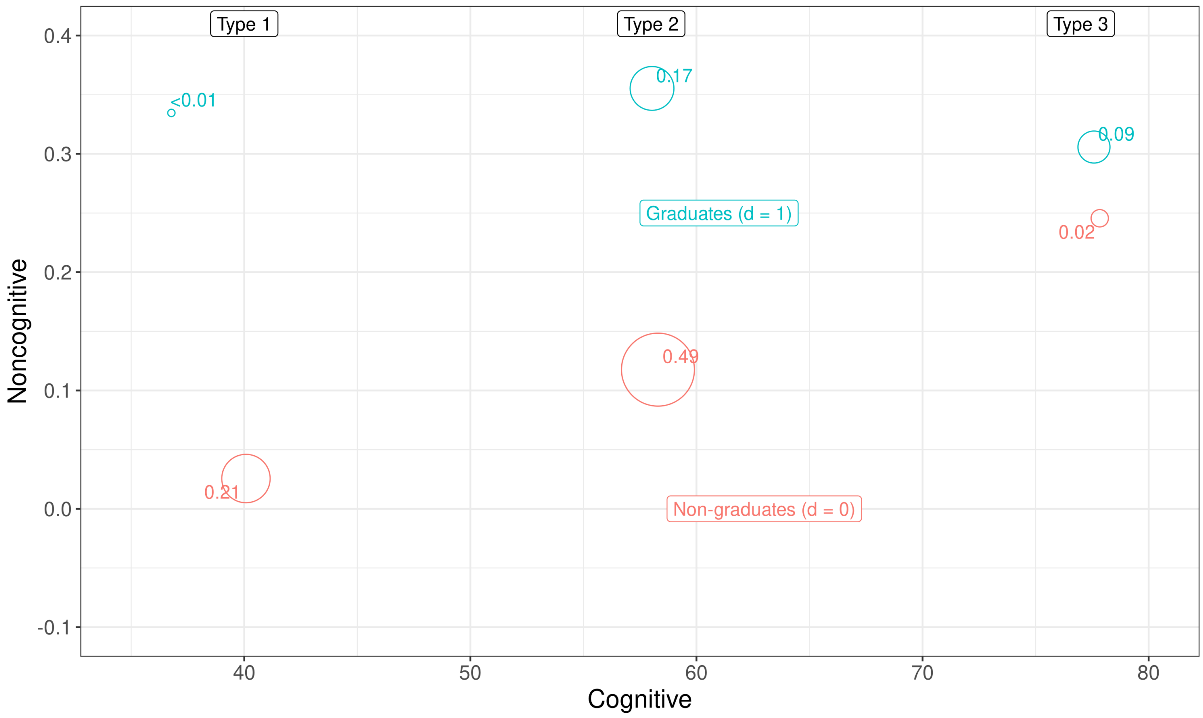
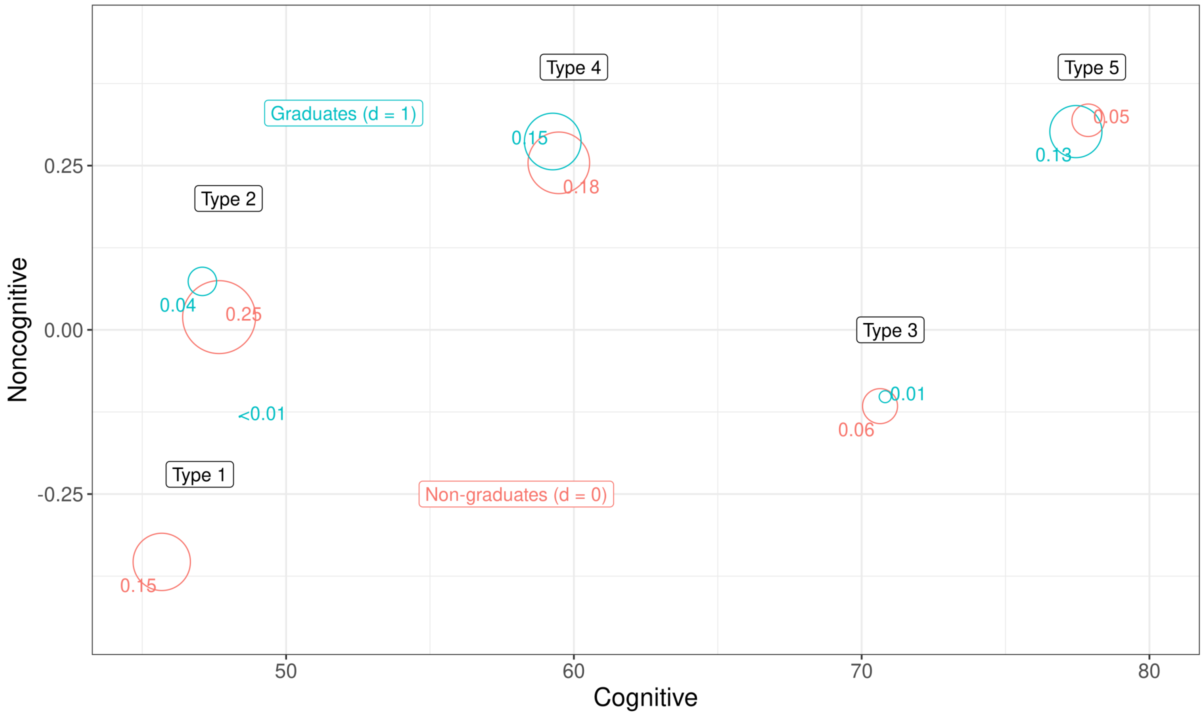
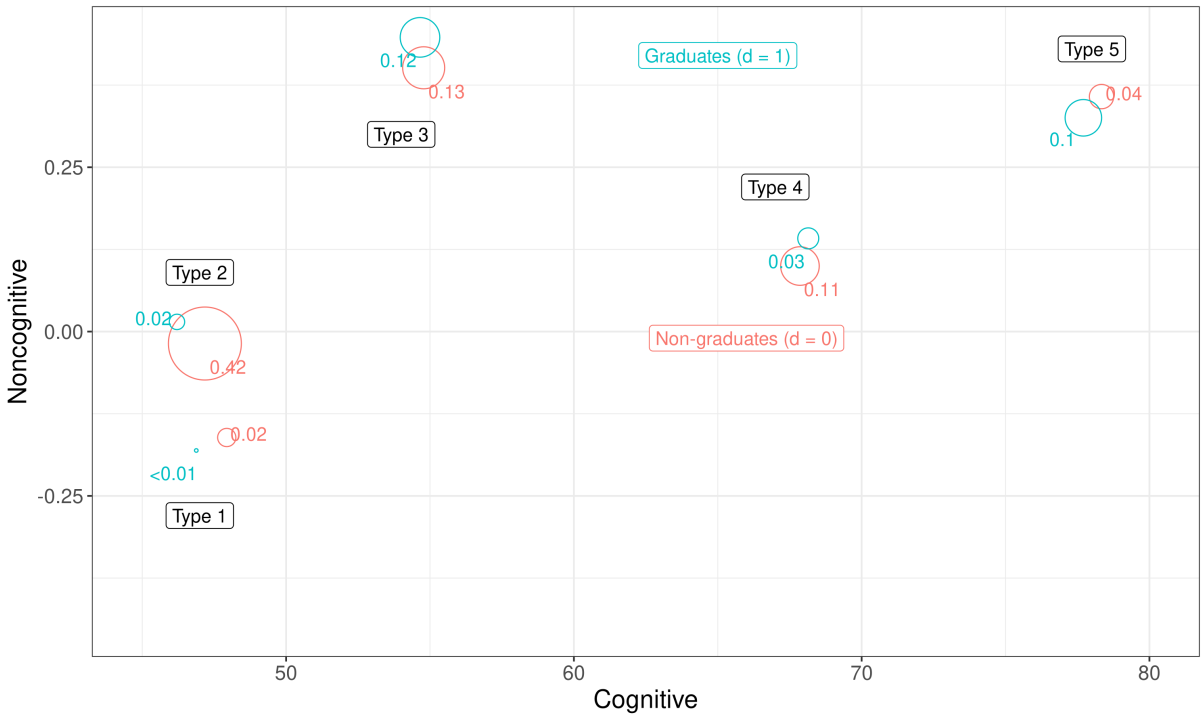


Figure 4: Types in cognitive-noncognitive space ($K = 5$, both measures)

(a) Male



(b) Female



4.3 Cognitive and noncognitive measures of human capital

The results obtained with both cognitive and non-cognitive measurements are in figure 4 and in table 5. Figure 4 plots the types on axes corresponding to cognitive and non-cognitive abilities, split into graduates (blue) and non-graduates (red), with the size of the hollow circles (and the labels) corresponding to the size of that group, $\pi(\theta = \theta_k)$. Including non-cognitive measures was successful in one sense at least; both cognitive and noncognitive abilities are now independent of education within types. We successfully match on both cognitive *and* non-cognitive skills, in contrast to figure 3. Once again we label types so that k is increasing in the mean wages of non-graduates, $\mu(\theta_k, 0)$.

We choose to present our estimates with five levels of human capital as that is the smallest number that captures key trends and allows variation in both components of human capital. Figure 5 displays, on the same axes as 4, the wages for graduates (proportional to the area of the hollow circles) and non-graduates (filled circles) and the wage premia for each type (white text). Some interesting patterns emerge. First, moving from type 1 to type 2 represents an increase in *non-cognitive* skills only; the cognitive skills of these types are very similar for both men and women. Type 2 for both men and women has a higher wage premium than type 1, despite these types having very similar non-graduate wages. The patterns for men and women now start to differ.

For men, the next type (3, as defined by non-graduate wages) has much higher cognitive skills, but *lower non-cognitive skills*. This type has the highest return to university. To get to type 4 we move north-west, representing a reduction in cognitive skill but a large increase in noncogntive skills. This type has a low return to university, though this appears to be driven by high *non-graduate* wages (particularly with respect to their level of cognitive skills). Finally, to reach type 5 we move (almost directly) east: a large increase in non-cognitive skills, and very similar non-cognitive skills. Type 5 has similar non-graduate wages to type 4, but higher graduate wages and hence a higher wage premium (22% vs 12%). Overall for men, it seems that non-cognitive skills are important for *non-graduate* wages, and cognitive skills are more important for *graduate* wages.

We turn our focus now to women, in panel (b) of figure 5. The cognitive and noncognitive skills of the types are reasonably similar across genders, though these abilities are renumerated quite differently for women compared to men. After type 2, the next type by non-graduate wage has slightly higher non-cognitive skills, and much higher cognitive skills. Type 3 has a similar non-graduate wage to type 2, but higher graduate wages resulting in a higher wage premium. Similar to our findings for men, type 4 is north-west of type 3 in "ability-space" for women, having higher non-cognitive skills and lower cog-

[14] Recall that under our model of (noisy) measures and outcomes, the type means are directly informative of human capital, although individual measures and outcomes are not.

nitive skills. Type 4 for women have the highest non-cognitive skills across all types and genders. Their graduate and non-graduate wages are very similar to those of type 3 (who have lower non-cognitive but higher cognitive skills), suggesting for women cognitive and non-cognitive skills are substitutes. Finally female type 5 has slightly lower non-cognitive skills relative to type 4, but higher cognitive skills. These are paid well (though still significantly less than similar ability men) as both graduates *and* non-graduates, resulting in a reduced wage premium relative to types 2, 3 and 4 (16% vs 22, 28, and 27% respectively).

Our results suggest that for men the occupations that graduates and non-graduates enter are quite different, with cognitive skills better rewarded in graduate occupations, while non-cognitive skills are (relatively) better rewarded in non-graduate occupations. The same cannot be said for women, for whom (prior) cognitive and non-cognitive skills are substitutable to a similar extent across graduate and non-graduate occupations.

Comparing the returns to human capital with the returns to university. A final analysis we can perform using the results in table 5 is to compare the effects of a low-ability individual graduating from university, with a (hypothetical) increase in their human capital. The low returns for type 1 of both genders make this uninteresting. However, an interesting comparison involves the wages of a type 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)[15] graduating from university than increasing their human capital to the level of the highest ability type. For women, type 2 would earn about the same in either counterfactual. This emphasises how high the wage returns are to a university, even for lower ability young people.

4.4 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 human capital. However, it is still interesting to place our results in the context of previous work. In section 2.4 we showed how we could 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 6, 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 calculated using the "classic" formula, and shown in table 3.

Our ATE and ATT estimates are similar to the OLS estimates on our data, and estimates

[15]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 5: Results by type with cognitive and noncognitive measures ($K = 5$)

(a) Male

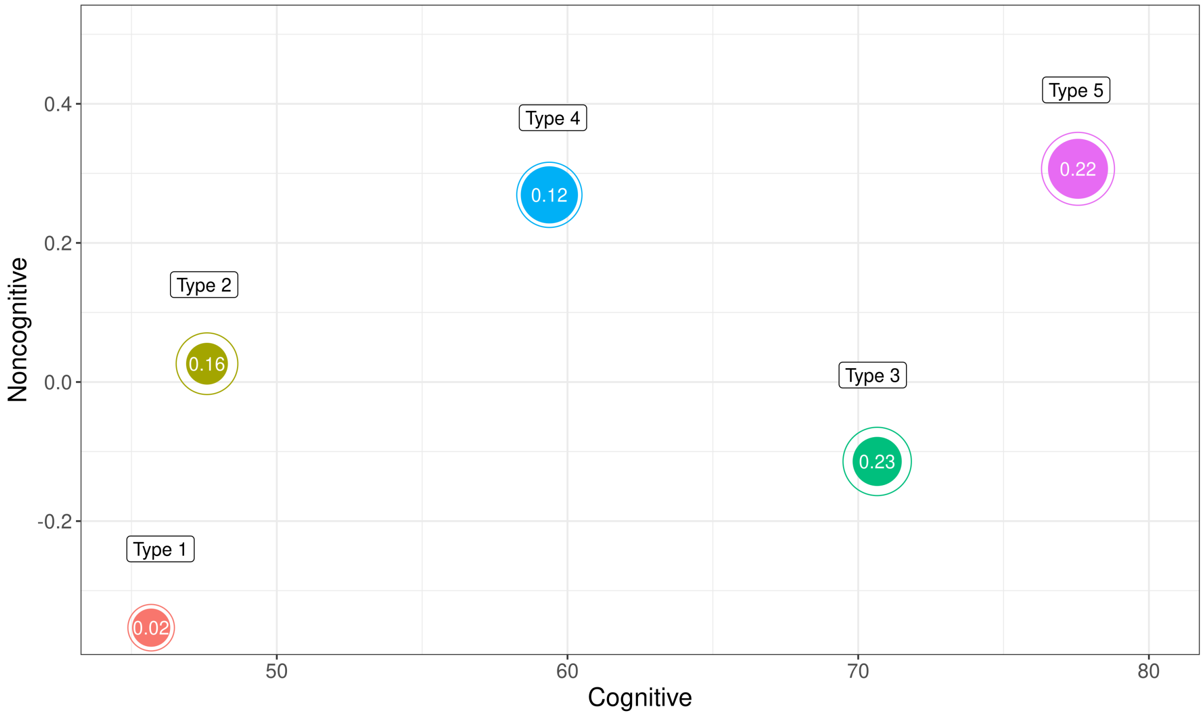
Type, $k =$	1		2		3		4		5	
Education, $d =$	0	1	0	1	0	1	0	1	0	1
Returns	0.024		0.157		0.228		0.115		0.216	
Wage (age 25, GBP)										
Mean	207	212	207	242	213	268	227	255	234	290
Ability measures										
Cognitive	45.7	48.4	47.7	47.1	70.6	70.8	59.5	59.3	77.9	77.4
Non-cognitive	-0.35	-0.13	0.02	0.07	-0.12	-0.10	0.25	0.29	0.32	0.30
$\pi(\theta_k)$	0.15	<0.01	0.25	0.04	0.06	0.01	0.18	0.15	0.05	0.13

(b) Female

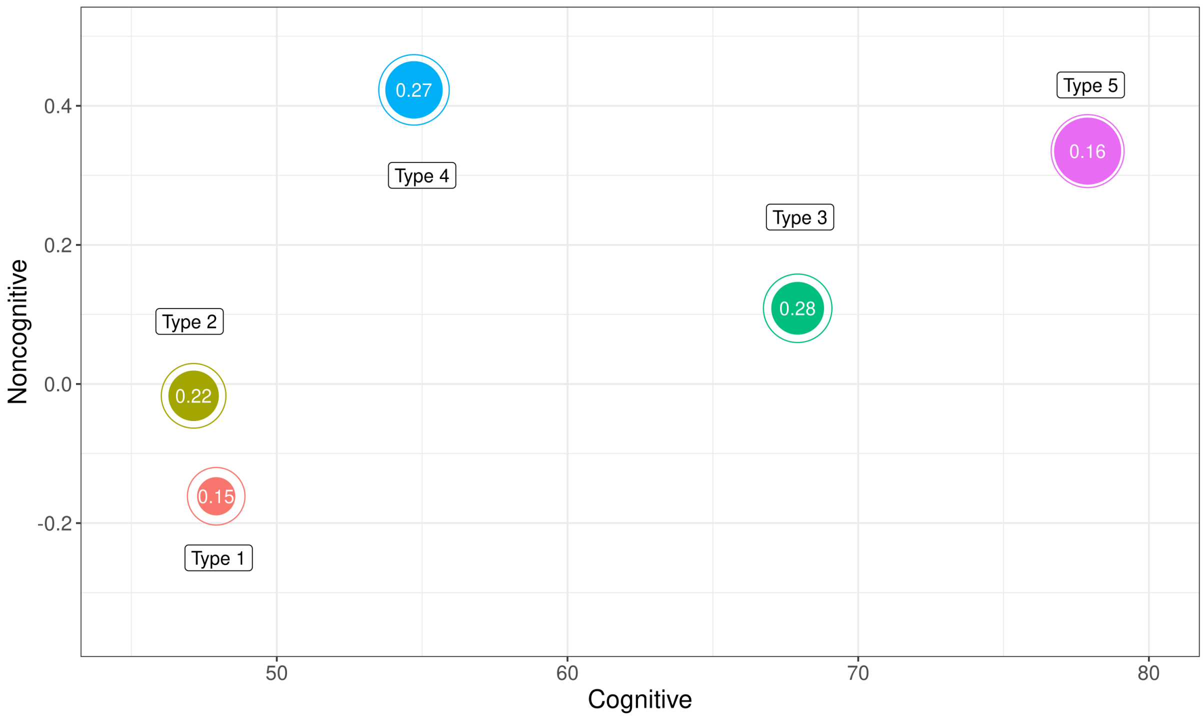
Type, $k =$	1		2		3		4		5	
Education, $d =$	0	1	0	1	0	1	0	1	0	1
Returns	0.150		0.219		0.278		0.267		0.164	
Wage (age 25, GBP)										
Mean	143	166	151	188	154	204	163	213	192	227
Ability measures										
Cognitive	47.9	46.9	47.2	46.2	67.9	68.1	54.8	54.7	78.3	77.7
Non-cognitive	-0.16	-0.18	-0.02	0.01	0.10	0.14	0.40	0.45	0.36	0.33
$\pi(\theta_k)$	0.02	<0.01	0.42	0.02	0.11	0.03	0.14	0.12	0.04	0.10

Figure 5: Estimated distribution of wage premia (cog-noncog space, $K = 5$)

(a) Male



(b) Female



Notes: Panel (a) 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 noted in white on each circle.

Table 6: 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

Note:

(b) Weights in bias formula

	Male					Female				
Type $k=$	1	2	3	4	5	1	2	3	4	5
Weights										
B_{OLS}	-0.224	-0.252	-0.063	0.212	0.328	-0.030	-0.509	-0.037	0.259	0.316

of other authors on UK data from a similar period.[16] Comparing our model estimates with the OLS and IV estimates using our data there are a number of things to note. First, the OLS estimates are broadly similar to the model estimates though are slightly biased relative to our ATT. In section 2.4 we derived a formula for the OLS estimator (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.

$$B_{OLS} = \sum_{\theta} \underbrace{[\pi(\theta \,|\, d=1) - \pi(\theta \,|\, d=0)]}_{\text{weights}} \mathbb{E}[w_0 \,|\, \theta]$$

Panel (b) of table 6 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: the positive and negative weights cancel out.

Correlation between wage premia and human capital. We perform an exercise to estimate the correlation between human capital and wage premia in our estimated model, designed to identify non-linearities in the returns. We have already established that the relationship between human capital, wages, and the returns to university is non-linear in our analysis by level of human capital. Here we provide further evidence of this non-linearity. We calculate for each individual their "true" cognitive and non-cognitive human

[16] In section 3.2 we described how 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.

Table 7: Correlations between wage premium and abilities ($K = 5$)

	Dependent variable: $w_1 - w_0$					
	Male			Female		
	(1)	(2)	(3)	(1)	(2)	(3)
Cognitive	0.037***	−0.045***	−0.099***	−0.021**	−0.118***	−0.102***
	(0.009)	(0.008)	(0.005)	(0.009)	(0.009)	(0.008)
Cognitive^2			0.224***			0.317***
			(0.007)			(0.010)
Non-cognitive	0.009	0.110***	0.117***	0.024*	0.131***	−0.150***
	(0.012)	(0.010)	(0.007)	(0.015)	(0.013)	(0.011)
Non-cognitive^2			0.146***			0.290***
			(0.008)			(0.012)
Cog×Non-cog		0.132***	−0.290***		0.253***	−0.600***
		(0.006)	(0.014)		(0.011)	(0.022)
Observations	745	745	745	1,131	1,131	1,131
Note:				*p<0.1; **p<0.05; ***p<0.01		

capital, and an expected wage premium.[17] We then estimate the following regression,

$$w_{1i} - w_{0i} = \beta_0 + \beta_C \theta_i^C + \beta_N \theta_i^N + \beta_{2C} (\theta_i^C)^2 + \beta_{2N} \theta_i^N + \beta_{CN} (\theta_i^C \cdot \theta_i^N) + \nu_i$$

where θ_i are the "true" cognitive and non-cognitive human capital that we assign to each individual, and $w_{1i} - w_{0i}$ is the deterministic wage premium. The results of this estimation are in table 7. Most of the coefficients are significantly different from zero at the 1% level, and in particular this is true of all the non-linear terms: $(\theta_i^C)^2$, $(\theta_i^N)^2$ and the interaction $\theta_i^C \cdot \theta_i^N$. This reinforces our claim that allowing for non-linearities is important.

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 ability;

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

[17] There are two ways to achieve this: (i) we can assign each individual to their closest type, given by the maximum of their posterior probabilities; or (ii) we can use the posterior probabilities to directly assign an "idiosyncratic" human capital vector and wage premia. Both methods are identical in the context of this exercise.

Table 8: 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

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

Formally, the decomposition is

$$\mathbb{V}(w) = \underbrace{\mathbb{E}[\mathbb{V}(\mathbb{E}[w\,|\,\theta, d]\,|\,d)]}_{\text{"within"}} + \underbrace{\mathbb{V}(\mathbb{E}[w\,|\,d])}_{\text{"between"}} + \underbrace{\mathbb{E}[\mathbb{V}(w\,|\,\theta, d)]}_{\text{"unexplained"}} \tag{11}$$

which allows us to compare the contributions of human capital before university and of the returns to university to wage inequality. The results are in table 8. 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 human capital 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 4.3 showing the wage gain from graduating for a low-ability young person (type 2) are equivalent to (hypothetically) becoming a graduate of the highest ability (type 5).

5 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 of this widely used model. Although we are not the first to show non-parametric identification, we achieve identification with fewer measurements of human capital than the current leading approaches in the literature. We are also the first to take the next step, estimating our model without imposing linearity in wages or measurements.

In our application we use 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 human capital 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 human capital on wages, and on the returns to a univeristy degree. The returns to university are also shown to be more important than the returns to 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 human capital 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 3 to 4 times larger than the contribution of ability.

References

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.

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

Blanden, J. and S. Machin (2004): "Educational Inequality and the Expansion of UK Higher Education," *Scottish Journal of Political Economy*, 51, 230–249.

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.

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.

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.

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," Working Paper 9546, National Bureau of Economic Research.

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.

Chetty, R., N. Hendren, P. Kline, E. Saez, and N. Turner (2014): "Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility," *American Economic Review*, 104, 141–147.

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.

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)*, 39, pp. 1–38, publisher: Wiley for the Royal Statistical Society.

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.

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,” .

Hu, Y. H. and M. Shum (2012): “Nonparametric identification of dynamic models with unobserved state variables,” *Journal of Econometrics*, 171, 32–44.

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.

Kasahara, H. and K. Shimotsu (2009): “Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices,” *Econometrica*, 77, 135–175.

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.

OECD (2019): *Education at a Glance 2019: OECD Indicators*, Education at a Glance, OECD.

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.

The Economist (2015): “The world is going to university,” *The Economist*.

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.

UNESCO (2020): “World Development Indicators,” .

A Nonparametric identification proof

Step 1: constructing matrices

- Likelihood of i's observations

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

- we can discretize these pdfs and store the resulting pmfs in matrices[18]

$$\begin{aligned}
\underset{n_{\boldsymbol{m}}\times n_w}{P(z,d)} &\equiv \Big[p(z,d,\boldsymbol{m},w)\Big]_{\boldsymbol{m}\times w}, \ \underset{K\times K}{D(z,d)} \equiv \operatorname{diag}\Big[\pi(\boldsymbol{k},z,d)\Big]_{\boldsymbol{k}\times\boldsymbol{k}} \\
\underset{n_{\boldsymbol{m}}\times K}{F_1} &\equiv \Big[f_m(\boldsymbol{m}|\boldsymbol{k})\Big]_{\boldsymbol{m}\times\boldsymbol{k}} \\
\underset{n_w\times K}{F_2} &\equiv \Big[f_w(w|\boldsymbol{k},d)\Big]_{w\times\boldsymbol{k}}, \ K = K_1 K_2
\end{aligned}$$

- Then for a given z, d

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

Assumption 1

(a) $\pi(\boldsymbol{k},z,d) \neq 0$ *all types have positive probability of (not) attending university* $\forall z$

(b) The matrices F_{1j} $\forall j = 1,..,J$ *and* F_2 *are linearly independent (i.e. full column rank): different types have different wage distributions*

(c) $\frac{\pi(\boldsymbol{k},1,d)}{\pi(\boldsymbol{k},0,d)} \neq \frac{\pi(\boldsymbol{k}',1,d)}{\pi(\boldsymbol{k}',0,d)}$ $\forall \boldsymbol{k} \neq \boldsymbol{k}'$: *different types are affected differently by the instrument*

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

- Then for a given d (which we now omit to simplify the notation)

$$\begin{aligned}
P(0) &= F_1 D(0) F_2^\top \\
P(1) &= F_1 D(1) F_2^\top
\end{aligned}$$

  as F_1 and F_2 are independent of z

- By assumption 1(b) F_1 and F_2 are full column rank

- Assumption 1(a) ensures $D(0)$ and $D(1)$ are invertible.

[18] If $f_m(\boldsymbol{m}|\boldsymbol{k}) = \prod_\ell f_\ell(m_\ell|\boldsymbol{k}_{(j)})$, then $F_1 = \otimes_\ell F_{1\ell}$ with

$$\underset{n_\ell\times K_{(j)}}{F_{1\ell}} \equiv \big[f_\ell(m_\ell|\boldsymbol{k}_{(j)})\big]_{m_\ell\times\boldsymbol{k}_{(j)}}$$

- Then $P(0)$ has SVD: $P(0) = U\Sigma V^\top$
- We can partition $U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}$ and $V = \begin{bmatrix} V_1 & V_2 \end{bmatrix}$ so that

$$P(0) = U_1 \Sigma_1 V_1^\top$$

  where Σ_1 contains the K non-zero singular values of $P(0)$ on its diagonal.
- Then

$$\begin{aligned} \Sigma^{-1} U^\top P(0) V &= I_K \\ \Leftrightarrow \underbrace{\Sigma^{-1} U^\top F_1}_{W} \underbrace{D(0) F_2^\top V}_{W^{-1}} &= I_K \end{aligned}$$

- and for $z = 1$

$$\begin{aligned} \Sigma^{-1} U^\top P(1) V &= \Sigma^{-1} U^\top F_1 D(1) F_2^\top V \\ &= \Sigma^{-1} U^\top F_1 D(1) D(0)^{-1} D(0) F_2^\top V \\ &= W D(1) D(0)^{-1} W^{-1} \end{aligned}$$

- The non-zero (diagonal) entries of

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

  are the eigenvalues of $\Sigma^{-1} U^\top P(1) V$
- and W must contain eigenvectors of $\Sigma^{-1} U^\top P(1) V$
- We can determine these eigenvectors by noticing the sum down each column of F_1 is equal to 1 (as these are probability distributions)
- Then $W = \Sigma^{-1} U^\top F_1$ and hence F_1 are identified
- Inverting W and noting the columns of F_2 are also prob distributions identifies $D(0)$ and F_2
- Finally $D(1)$ is identified from $D(1)D(0)^{-1}$
- We can recover F_{1j} by summing within and across blocks of $F_1 = \bigotimes_j F_{1j}$

Correct labels across d

- To ensure correct labelling across d notice F_1 is independent of d
- These arguments extend to multiple treatments (i.e. multivalued d)

B EM algorithm details

E-step.

In the E-step we update the posterior type probabilities, $p_i(k|\Theta)$:

$$p_i\left(\boldsymbol{k}|\hat{\Theta}^{(s)}\right) \equiv \frac{\hat{p}_k^{(s)}\ell\left(\hat{\Theta}^{(s)};\boldsymbol{m}_i,w_i,z_i,d_i,k\right)}{\sum_{k=1}^{K}\hat{p}_k^{(s)}\ell\left(\hat{\Theta}^{(s)};\boldsymbol{m}_i,w_i,z_i,d_i,k\right)}. \tag{12}$$

M-step.

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

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

$$\alpha_j(\boldsymbol{k})^{(s+1)} \equiv \frac{\sum_i p_i\left(\boldsymbol{k}|\hat{\Theta}^{(s)}\right)m_{ij}}{\sum_i p_i\left(\boldsymbol{k}|\hat{\Theta}^{(s)}\right)} \tag{13}$$

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

$$\omega_j(\boldsymbol{k})^{(s+1)} \equiv \sqrt{\frac{1}{N}\sum_{i=1}^{N}p_i\left(\boldsymbol{k}|\hat{\Theta}^{(s)}\right)\left(m_{ij}-\alpha_j(\boldsymbol{k})^{(s+1)}\right)^2} \tag{14}$$

- Update $\mu(\boldsymbol{k},d)$, $\sigma(\boldsymbol{k},d)$.
  1. Again use weighted means, with weights $p_i(\boldsymbol{k}|\hat{\Theta}^{(s)})$ to update μ_d:

$$\mu_d^{(s+1)}(\boldsymbol{k}) \equiv \frac{\sum_{i:d_i=d}p_i\left(\boldsymbol{k}|\hat{\Theta}^{(s)}\right)w_i}{\sum_{i:d_i=d}p_i\left(\boldsymbol{k}|\hat{\Theta}^{(s)}\right)} \tag{15}$$

  2. And use the updated μ_d to update $\sigma_d(\boldsymbol{k})$:

$$\sigma_d^{(s+1)}(\boldsymbol{k}) \equiv \sqrt{\frac{\sum_{i:d_i=d}p_i\left(\boldsymbol{k}|\hat{\Theta}^{(s)}\right)\left(w_i-\mu_d^{(s+1)}(\boldsymbol{k})\right)^2}{\sum_{i:d_i=d}p_i\left(\boldsymbol{k}|\hat{\Theta}^{(s)}\right)}} \tag{16}$$

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

$$\pi^{(s+1)}(\boldsymbol{k}, z, d) \equiv \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in I(z,d)} p_i(k|\hat{\Theta}^{(s)}), \tag{17}$$

  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:

$$\mathcal{L}(\boldsymbol{\Theta}^{(s)}; \boldsymbol{M}, \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{d}) - \mathcal{L}(\boldsymbol{\Theta}^{(s-1)}; \boldsymbol{M}, \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{d}) < \delta, \tag{18}$$

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

C Data

D Results

D.1 Choosing K

D.2 Single cognitive measure

D.3 Aggregate results

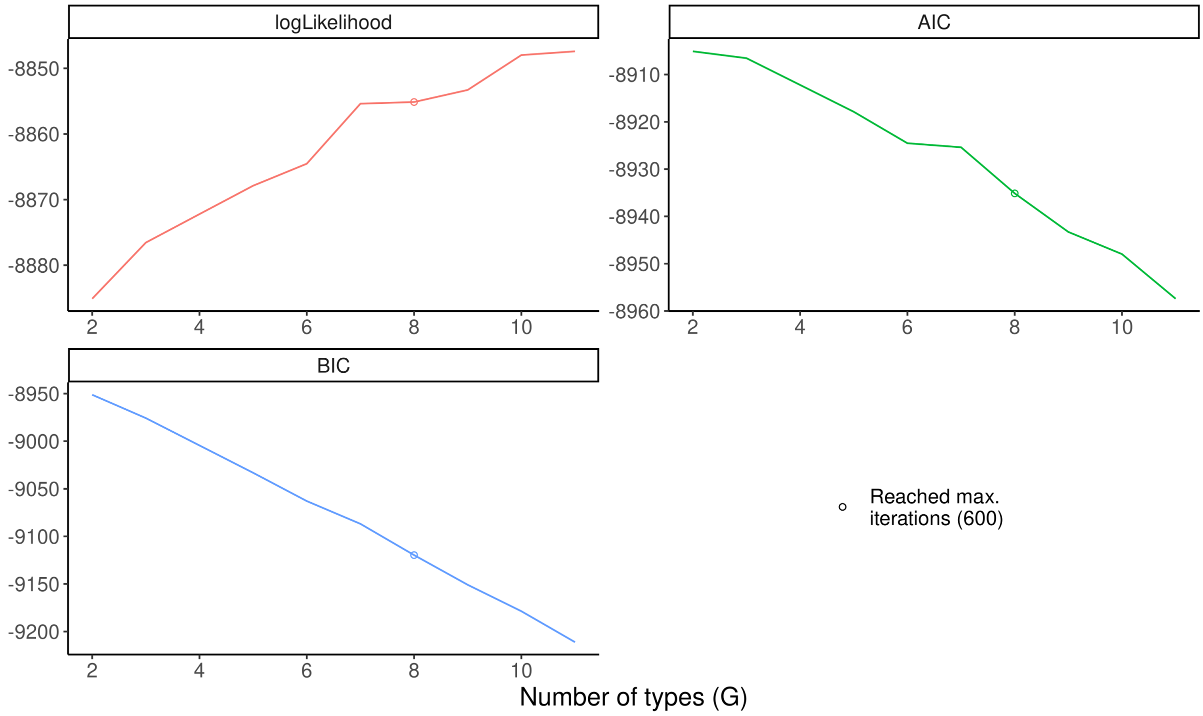
Table C9: Balancing checks for instrument validity: regions

	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.

Figure D6: Likelihood criteria: single cognitive measure

(a) Male



(b) Female

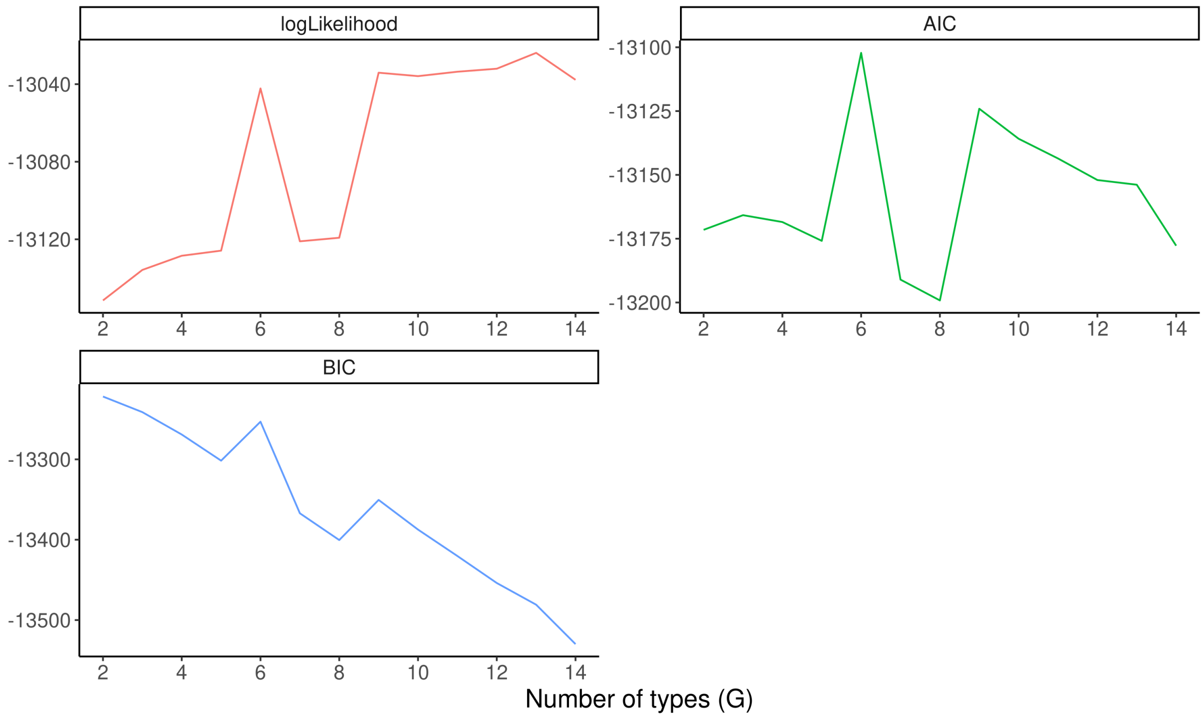
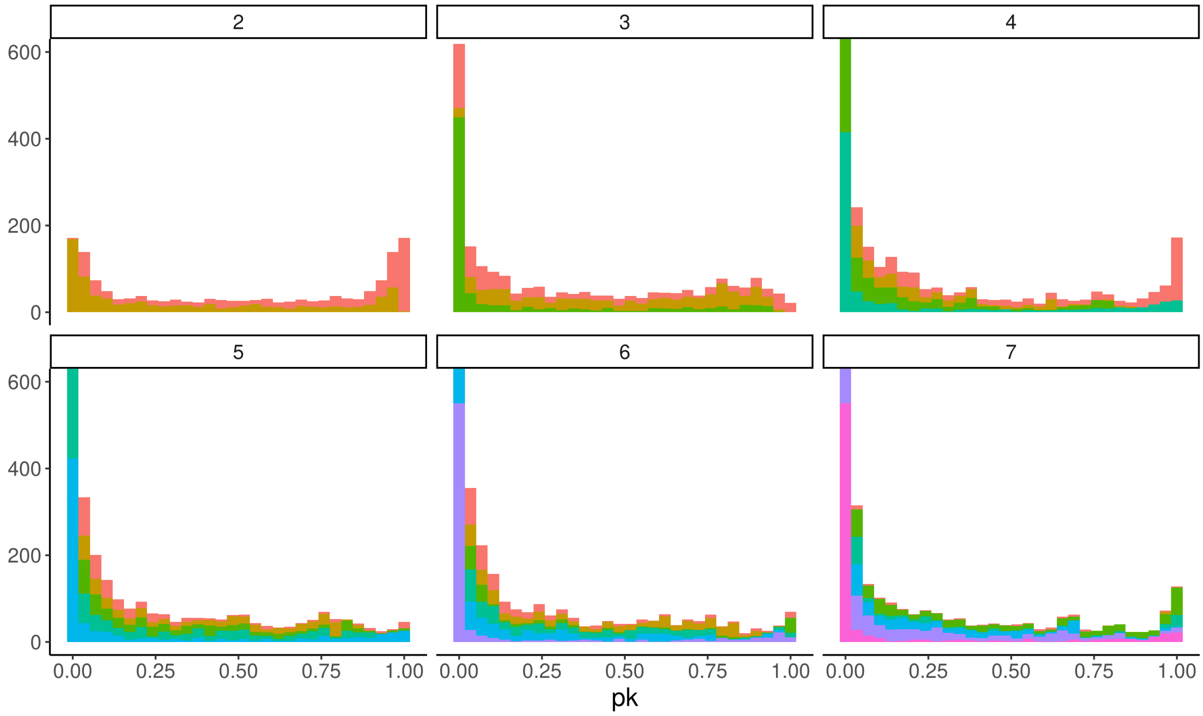


Figure D7: Posterior probabilities: single cognitive measure

(a) Male



(b) Female

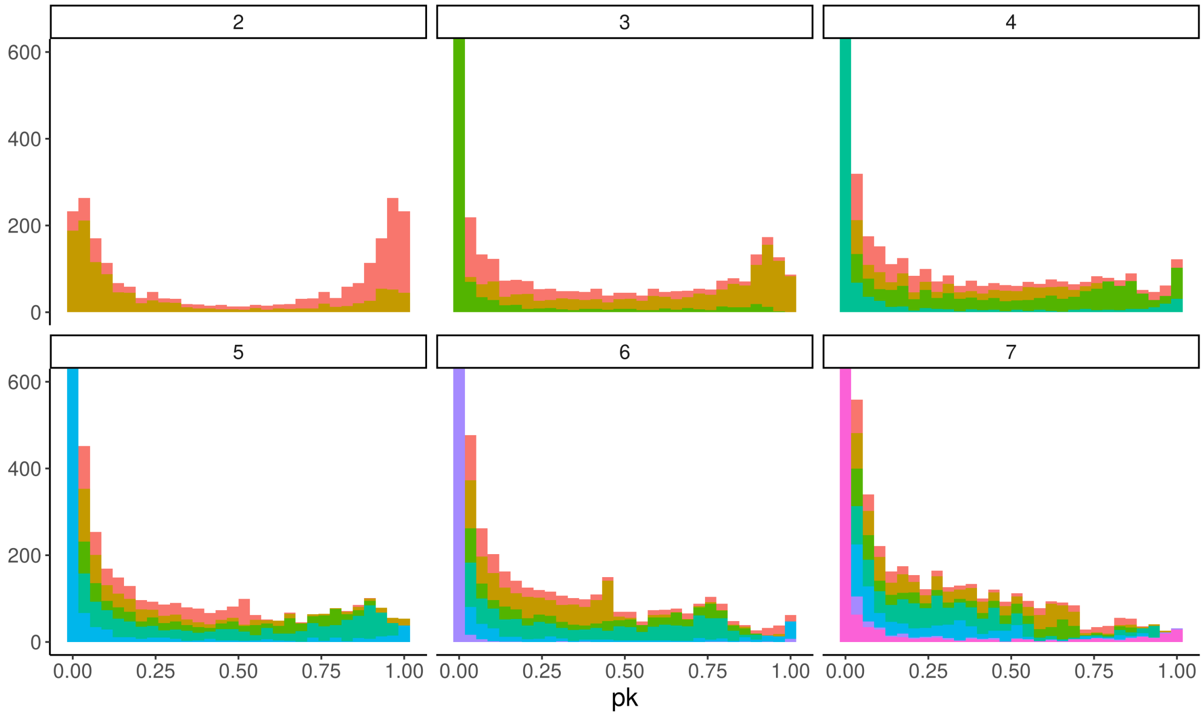
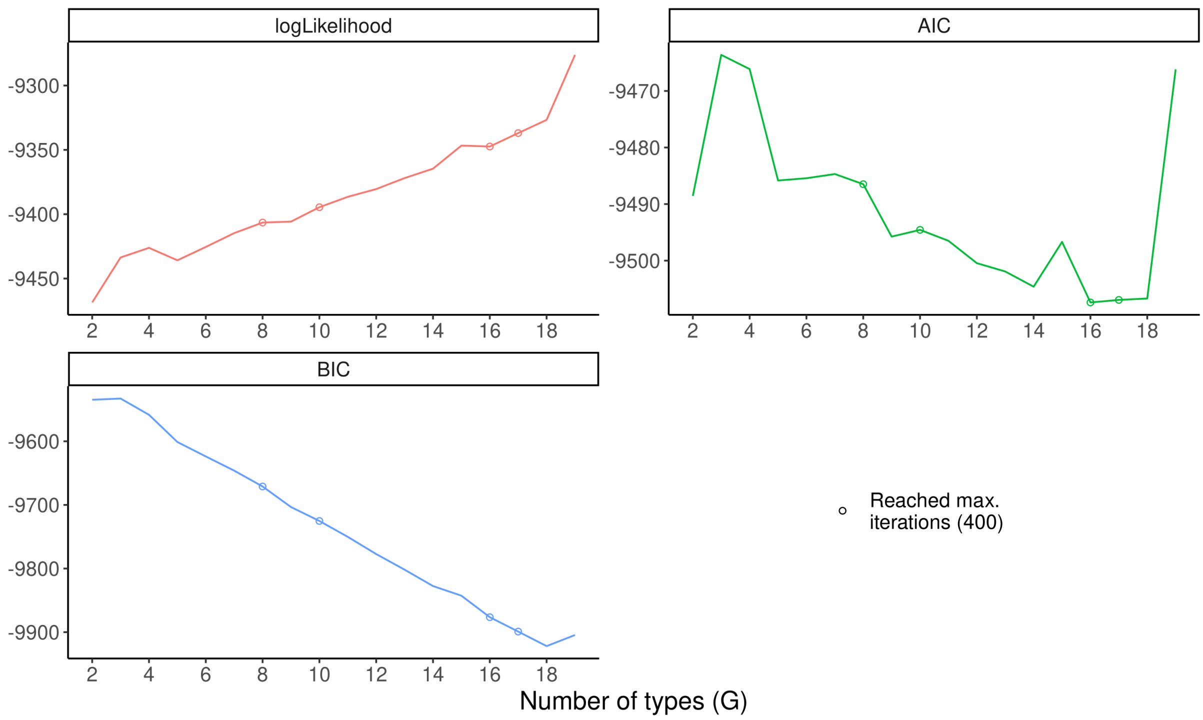


Figure D8: Likelihood criteria: cognitive and non-cognitive measures

(a) Male



(b) Female

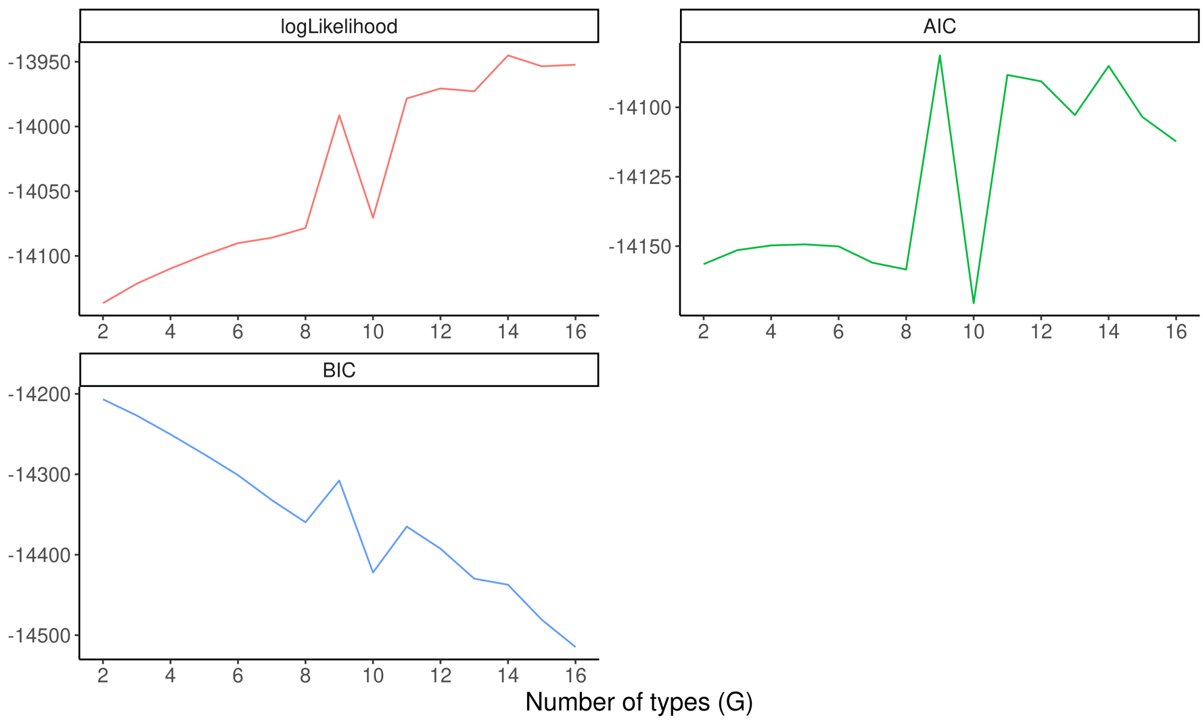
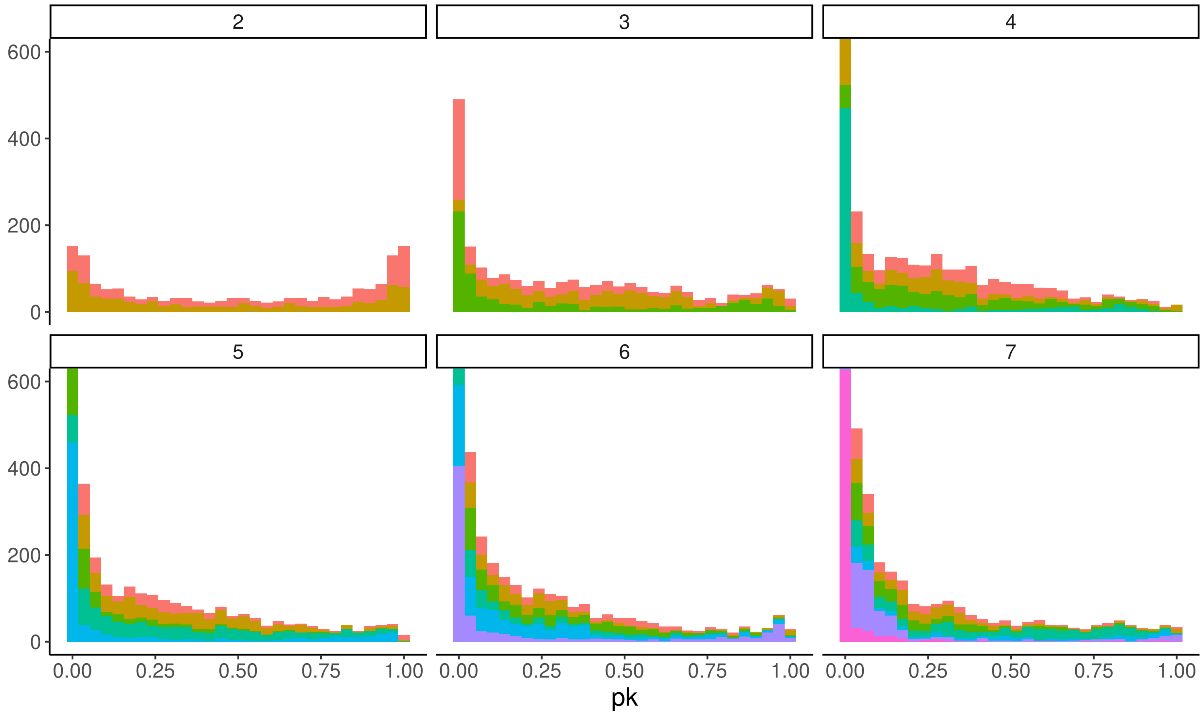


Figure D9: Posterior probabilities: cognitive and non-cognitive measures

(a) Male



(b) Female

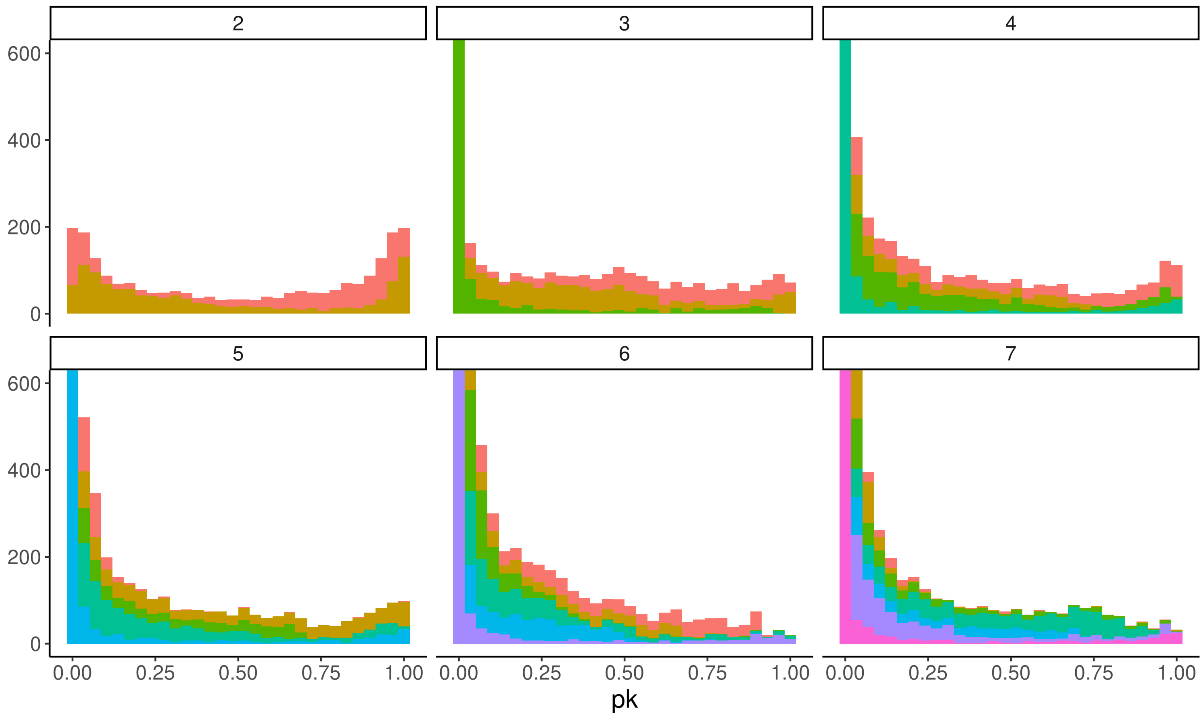
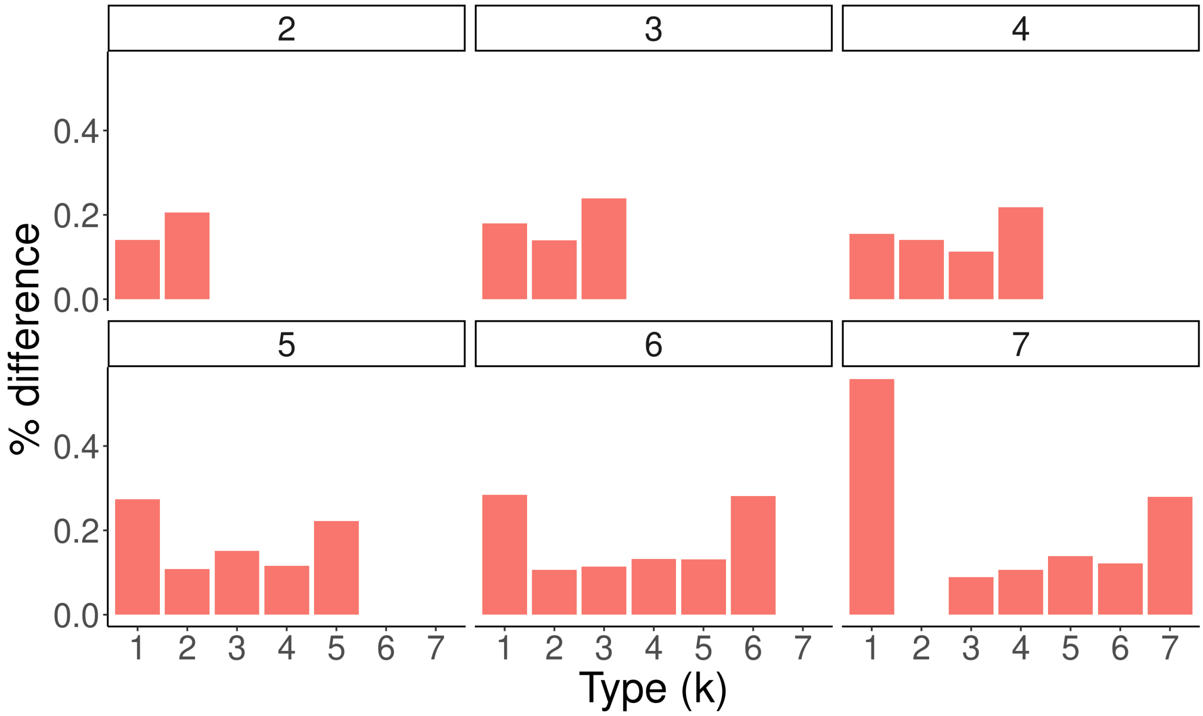


Figure D10: Results across K: single cognitive measure

(a) Male



(b) Female

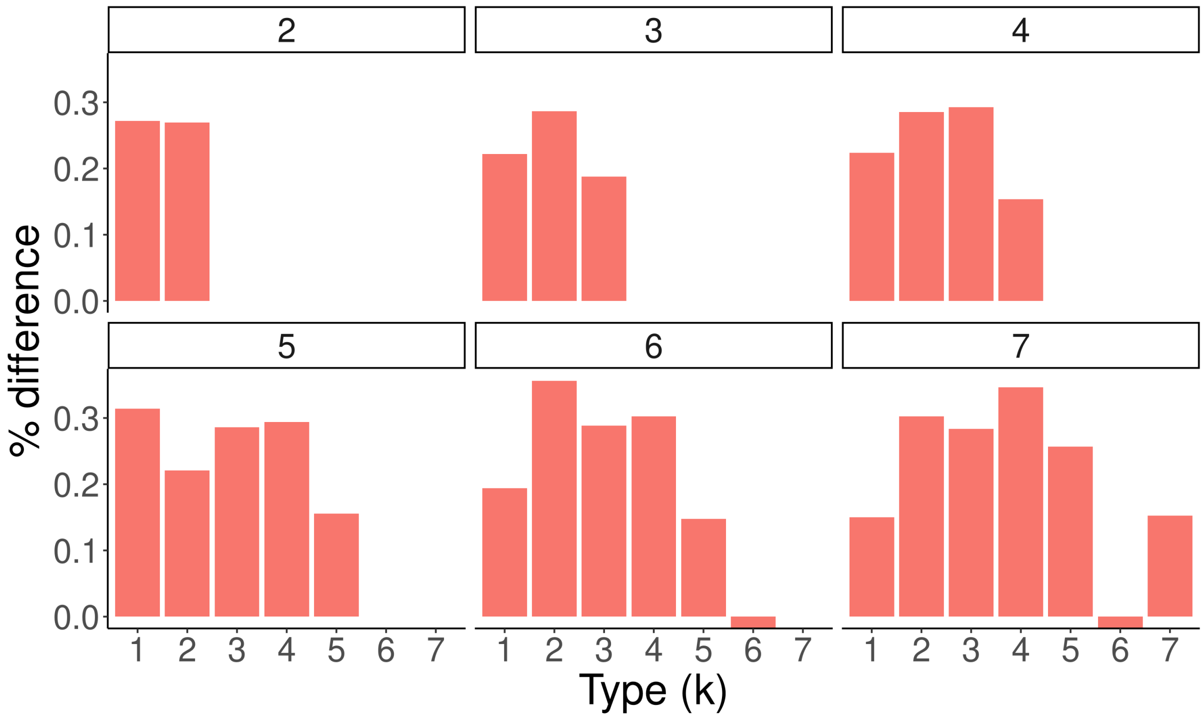
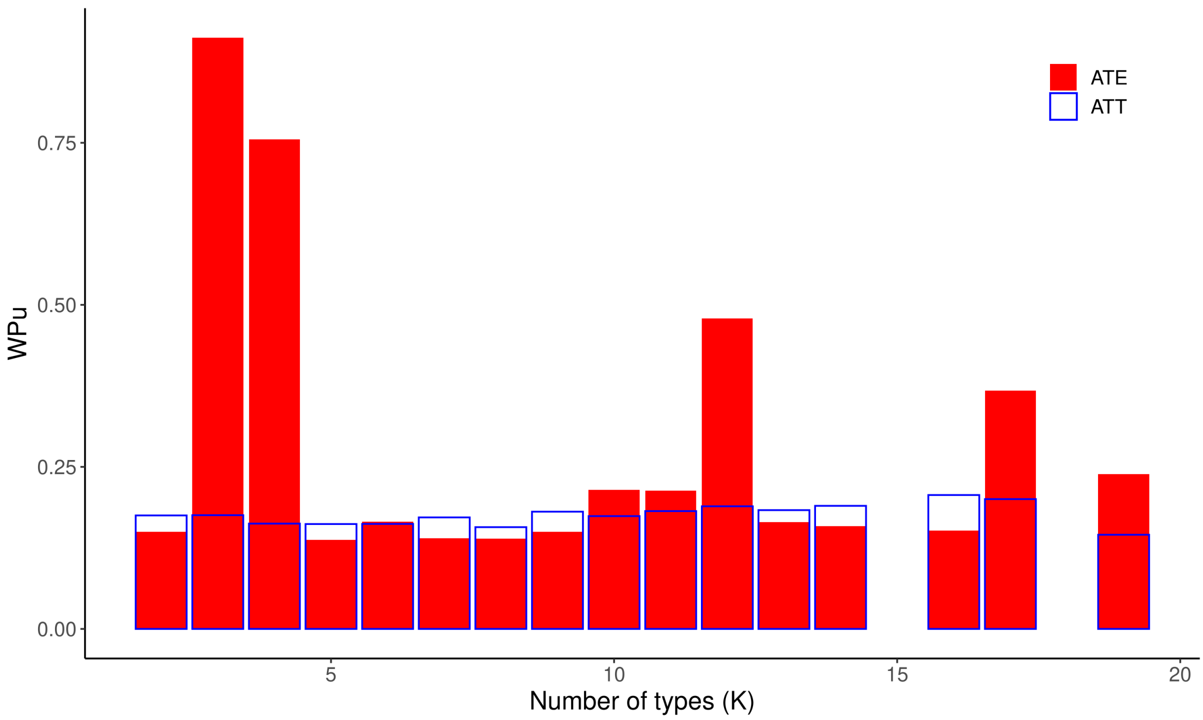


Figure D11: ATEs / ATTs across K: cognitive and noncognitive measures

(a) Male



(b) Female