Non-pecuniary returns to higher education: The effect on smoking intensity in the UK

Massimiliano Bratti[*], **Alfonso Miranda**[†‡]

Abstract. This paper investigates whether higher education (HE) produces non-pecuniary returns via a reduction in the intensity of consumption of health-damaging substances. In particular, it focuses on current smoking intensity of the British individuals sampled in the 29-year follow-up survey of the 1970 British Cohort Study. We estimate endogenous dummy ordinal response models for cigarette consumption and show that HE is endogenous with respect to smoking intensity and that even when endogeneity is accounted for, HE is found to have a strong negative effect on smoking intensity. Moreover, pecuniary channels, such as occupation and income, mediate only a minor part of the effect of HE. Our results are robust to modelling individual self-selection into current smoking participation (at age 29) and to estimating a dynamic model in which past smoking levels affect current smoking levels.

JEL classification: C35, I12, I21.

Keywords: endogenous dummy, ordinal response models, higher education, sample selection, smoking, UK.

---

[*]Department of Economics, Business and Statistics, University of Milan, Via Conservatorio 7, I-20122, Milan, Italy. E-mail: Massimiliano.Bratti@unimi.it

[†]Department of Quantitative Social Science, Institute of Education, University of London. 20 Bedford Way, London WC1H 0AL, UK. E-mail: A.Miranda@ioe.ac.uk

[‡]Both authors equally contributed to this paper. We acknowledge the original data depositor, the UK Data Archive (UKDA), which made available data from the 1970 British Cohort Study to us, and the original data collectors, the Department of Child Health at Bristol University and the Centre for Longitudinal Studies at the Institute of Education (University of London). Neither the original data collectors nor the data depositor bear any responsibility for the analyses or interpretations presented here. We are also grateful to Andrew Clark, Paul Clarke, Francesca Cornaglia, Gauthier Lanot, Martin Forster, Leslie Rosenthal and participants to presentations held at the University of Essex (ISER), the University of Warwick, EALE 2006 (Prague) and LACEA-LAMES 2006 (Mexico City) for useful suggestions. We are indebted to two anonymous referees and the Associate Editor Jennifer Roberts whose comments greatly improved our paper. The usual disclaimers apply.

1 Introduction

Higher education (HE) has been found to generate significant wage returns in the UK (**??**). Besides market returns, there is general agreement that education also has important non-pecuniary returns (see, for instance, the survey in **?**). As for health, for example, it has been shown that education has causal effects on individuals' health-related behaviour such as smoking. The existing economic literature has mainly focused on the effect of education on the extensive margin, that is on smoking participation (see **????**, among others), while evidence on the intensive margin is much scarcer. However, in the medical literature, smoking intensity has been shown to be related to a number of health problems and diseases. Just to take a few examples, women who smoke heavily have reduced fertility (**?**) and a higher risk of cervical cancer (**?**) and heavy smoking is also associated with a higher risk of stroke (**?**). The health risks of both active and passive smoking are well documented and the UK government has been recently active in reducing the diffusion of tobacco use.

For these reasons, the central focus of our paper will be on the effect of higher education on the smoking intensive margin, that is, we will investigate the links between higher education and the quantity of cigarettes smoked daily in the UK. Our study benefits from a unique longitudinal data set, the 1970 British Cohort Study (BCS70) rich in family background information on the sampled individuals.

We make a contribution to the existing literature in several respects. First, unlike most of the previous studies, we model the smoking intensive margin. Second, we explicitly take into account the nature of the self-reported smoking data, which are characterized by clustering of answers around multiples of five, using an ordinal response model instead of a count model. Third, we address the potential endogeneity of higher education with respect to smoking intensity using an Endogenous Dummy Ordered Probit (ED-OP) model and we also assess the robustness of our results with respect to both non-random selection into smoking participation and to making the

model dynamic.

The structure of the paper is as follows. Section 2 describes the BCS70 data. Section 3 introduces the econometric model, the identification strategy and presents the main results. The last section summarizes our main findings and concludes.

2 Data

We use data drawn from the BCS70. The BCS70 began in 1970 when data were collected on the births and families of 17,198 babies born in England, Wales, Scotland and Northern Ireland from the 5^{th} to the 11^{th} of April. There are currently six complete follow-up surveys available: 5-year, 10-year, 16-year, 26-year, 29-year and 34-year. As to the variables included in our empirical analysis, data on smoking habits and the highest educational qualification were collected in the 29-year follow-up survey while all the other contextual variables were provided in the 10-year follow-up survey. In our main analysis, we use therefore a cross-section of individuals observed in the 29-year follow-up survey but who also answered to the 10-year follow-up survey.

The BCS70 was affected by some attrition.[1]The specific choice of the 10-year and 29-year follow-up surveys is motivated by the lower level of attrition recorded for these waves and the better quality of the data. Indeed, the size of the six complete follow-up surveys were 13,135 (5-year), 14,875 (10-year), 11,615 (16-year), 9,003 (26-year), 11,261 (29-year) and 9,665 (34-year) individuals, respectively.[2]In addition to having a lower attrition rate, the 10-year follow-up survey was also less affected by item non response compared to the 16-year wave. For instance, the 16-year follow-up survey has very

[1]The **?**) writes "Analysis of differential response comparing achieved samples and target samples for any follow-up, using data gathered during the birth and earlier follow-ups, show that the achieved sample are broadly representative of the target sample. However, as in other surveys, some groups (e.g. those from minority ethnic, low social class, and atypical family backgrounds) are under-represented" (p. 11).

[2]These are the numbers of records in the public released files of microdata.

poor information on smoking behaviour, which is available only for 52.8% of individuals in the wave.

Hence, because of the poor quality of smoking information in some waves and panel attrition, we prefer to focus the core of our analysis on smoking at age 29 and do not fully exploit the longitudinal nature of BCS70. It must be noted that since there is some evidence showing that education positively affects the likelihood of smoking cessation (see, for instance, **?**), focusing only on individuals at age 29 means that our estimates of the causal effect of HE on cigarette consumption might not provide the complete picture of the role of HE on smoking over an individual's life-cycle. However, we make an attempt to address this potential weakness in our analysis by: 1) assessing the robustness of our results when account is taken of individual self-selection into current smoking participation at age 29 (Section 3.3); 2) estimating smoking dynamics between age 29 and age 34 using also the 34-year BCS70 follow-up survey (Section 3.4).

The age 29 follow-up survey reports some information on smoking behaviour. In particular, the relevant questions for our purposes are:

> *Would you say that: (a) you've never smoked cigarettes; (b) you used to smoke cigarettes but not at all now; (c) you now smoke cigarettes occasionally but not every day; (d) you smoke cigarettes every day?*

and

> *How many cigarettes a day do you usually smoke?*

The distribution of valid answers to the first question is shown in Table 1. About 29% of individuals in our sample smoke cigarettes every day, while 7.7% smoke occasionally. Every-day smokers are asked the number of cigarettes smoked a day (i.e. the second question). In the present paper we consider all current smokers and set the number of cigarettes smoked a day at zero for occasional smokers.

[Table 1 about here]

Self-reported data on daily cigarette consumption may contain errors. In particular, when looking at the distribution of the number of cigarettes smoked daily S_i (see Figure 1) we observe that multiples of five are more likely to occur. This is a common phenomenon found in studies of cigarette consumption (see, for instance **?**, who use BHPS data).

In general, it is hard to say to what extent the peculiar pattern observed for cigarette consumption reflects a true pattern or is affected by measurement error. Our personal view is that current smokers, when reporting the number of cigarettes smoked a day, may approximate it to the closest multiple of five. In order to address this feature of the data we decide to discretise the number of cigarettes smoked into 5-cigarette bands. Hence, the dependent ordinal variable measuring smoking intensity (S_i^o) takes on the following values:

$$S_i^o = \begin{cases} 1 & \text{if } S_i = 0 \text{ (occasional smoker)} \\ 2 & \text{if } 5 \geq S_i > 0 \\ 3 & \text{if } 10 \geq S_i > 5 \\ 4 & \text{if } 15 \geq S_i > 10 \\ 5 & \text{if } 20 \geq S_i > 15 \\ 6 & \text{if } S_i > 20. \end{cases} \tag{1}$$

[Figure 1 about here]

Table 2 reports the average number of cigarettes smoked a day by level of education. HE refers to individuals with a degree level qualification or more. In particular, HE includes individuals with levels of qualification 4 or 5 in the standard UK classification of educational qualifications (see, for instance, **?**, p. 45). Male smokers with HE smoke about four cigarettes fewer a day than those with less than HE (-41%). The difference in the number of cigarettes smoked a day between female smokers with HE and those with less than HE is similar and amounts to almost four cigarettes (-48%).

[Table 2 about here]

In this paper we compare daily cigarette consumption of individuals with HE with those who have lower educational qualifications. Hence, individuals who achieved a HE degree are the "treatment group" while individuals with lower levels of education are the "control group". The composition of the control group (lower than HE) by highest educational qualification achieved is reported in Table 3. Table 3 clearly shows that the control group is mostly composed of individuals who did not go into post-compulsory schooling (79% for males and 85% for females). Hence, the effect of HE that we estimate can be roughly interpreted as the effect with respect to individuals with no more than compulsory education.

[Table 3 about here]

3 Empirical analysis

This section outlines the main features of our econometric model, discusses the model identification, presents the main empirical results and reports some additional findings.

3.1 Econometric model

In the empirical analysis we use an endogenous dummy model for ordinal response variables (see **?**). Denote by S_i^o the ordinal variable of interest for the $i-th$ individual. Variable S_i^o takes on a limited number of response categories S_{ih}^o, $h = 1, 2, \ldots, H$. Such categories are ordered, $S_{i1}^o < S_{i2}^o < \cdots < S_{iH}^o$, and the difference between any pair, $S_{ij}^o - S_{ih}^o$, does not necessarily accept a cardinal interpretation for all j and h. S_i^o is generated according to a continuous latent variable (smoking intensity equation),

$$S_i^* = \mathbf{x_i'}\boldsymbol{\beta} + v_i, \tag{2}$$

where $\mathbf{x_i}$ represents a $K \times 1$ vector of individual characteristics (excluding the constant term), $\boldsymbol{\beta}$ represents a conformable vector of coefficients, and v_i

represents a random error. The observed response, S_i^o, is determined by a threshold model,

$$S_i^o = \begin{cases} 1 & \text{if} \quad S_i^* \leq k_1 \\ 2 & \text{if} \quad k_1 < S_i^* \leq k_2 \\ . & . \quad . \\ . & . \quad . \\ H & \text{if} \quad k_{H-1} < S_i^* \end{cases}$$

where $\{k_1, \cdots, k_{H-1}\} \in \mathbb{R}^{H-1}$ are constants to be estimated along with other parameters in the model.

An important feature of the model is the fact that $\mathbf{x_i}$ includes an endogenous dummy, HE_i, which takes on value one if the i-th individual completed her undergraduate studies by the time of the survey and zero otherwise. The endogenous dummy HE_i is also generated by a latent variable model (HE equation)

$$HE_i^* = \mathbf{z_i'}\boldsymbol{\gamma} + w_i, \tag{3}$$

with

$$HE_i = \begin{cases} 1 & \text{if} \quad HE_i^* > 0 \\ 0 & \text{otherwise.} \end{cases}$$

As usual, $\mathbf{z_i}$ and $\boldsymbol{\gamma}$ are both $M \times 1$ vectors and w_i is a random error.

Correlation between v_i and w_i is induced by an unobserved heterogeneity term u_i that affects smoking and schooling decisions. In particular it is supposed that

$$\begin{aligned} v_i &= \lambda u_i + \eta_i \\ w_i &= u_i + \zeta_i, \end{aligned} \tag{4}$$

where $\lambda \in \mathbb{R}$ is a free parameter (*factor loading*) that is estimated within the model. We assume that u_i, η_i, and ζ_i are all independent standard normal variables and we label the model which follows the Endogenous Dummy Ordered Probit (ED-OP) model.

Define a set of dummy variables d_{ih} that take on value one if $S_i^o = h$ and zero otherwise, $h = \{1, \ldots, H\}$. The contribution of the i-th individual to

the likelihood is then written as,

$$L_i = \int_{-\infty}^{\infty} \sum_{h=1}^{H} d_{ih} \Phi_h^* \left\{ HE_i \Phi \left(\mathbf{z_i'} \boldsymbol{\gamma} + u_i \right) + \left(1 - HE_i \right) \left[1 - \Phi \left(\mathbf{z_i'} \boldsymbol{\gamma} + u_i \right) \right] \right\} \phi(u_i) du_i \tag{5}$$

with,

$$\Phi_h^* = \begin{cases} 1 - \Phi \left(\mathbf{x_i'} \boldsymbol{\beta} - \kappa_1 + \lambda u_i \right) \text{ if } h = 1 \\ \Phi \left(\mathbf{x_i'} \boldsymbol{\beta} - \kappa_{h-1} + \lambda u_i \right) - \Phi \left(\mathbf{x_i'} \boldsymbol{\beta} - \kappa_h + \lambda u_i \right) \text{ if } 1 < h \leq H - 1 \\ \Phi \left(\mathbf{x_i'} \boldsymbol{\beta} - \kappa_{H-1} + \lambda u_i \right) \text{ if } h = H \end{cases}$$

where $\Phi(\cdot)$ is the standard normal distribution function. The model is estimated by Maximum Simulated Likelihood (see, for instance, **?**).

Notice that, without loss of generality, specifying v_i and w_i as in (4) reduces the order of integration in (5) from two to one.

The correlation coefficient between v_i and w_i, ρ, and the factor loading λ are related in the following way:

$$\rho = \frac{\lambda}{\sqrt{2 \left(1 + \lambda^2 \right)}}.$$

One should be aware that the model with exogenous HE is nested within the endogenous dummy framework. This is so because if $\rho = 0$ the random terms w_i and v_i are independent and the likelihoods for the ordered variable and the endogenous dummy are separable — which is what exogenous dummy means in the econometrics literature (see **?**). This implies then that a test for the endogeneity of HE_i in equation (2) can be performed on the basis of a simple likelihood ratio test for $\rho = 0$.

Our econometric model will enable us to distinguish between some alternative hypotheses on the effect of HE on smoking. In particular, we will be able to distinguish between these four different situations: 1) the correlation coefficient ρ is not statistically different from zero and the coefficient on HE in the smoking equation is statistically significant. In this case HE is exogenous with respect to smoking behaviour and its effect is causal; 2) the correlation coefficient ρ is statistically significant while the coefficient

on HE in the smoking equation is not. In this case HE is endogenous and the correlation between HE and smoking behaviour is driven by unobserved heterogeneity (the so-called *third variable hypothesis* emphasised by **?**); 3) both the correlation coefficient ρ and the coefficient on HE in the smoking equation are significant. In this case although HE is endogenous with respect to smoking, it also has a causal impact on smoking behaviour. The estimates of ρ and of the causal effect of HE will also give an idea of the relative importance of the two alternative explanations, i.e. *third variable hypothesis* vs. causal effects; 4) the correlation coefficient and the coefficient on HE in the smoking equation are both insignificant. In this case our analysis will not support any of the hypotheses put forward above.

3.2 Empirical strategy and identification

The ED-OP model is formally identified through functional form (see **?**) and exclusion restrictions are unnecessary. However, although the model is formally identified, it may suffer from 'tenuous' identification and it may be useful to improve identification through some exclusion restrictions.[3]

The covariates included in the model were selected using a general-to-specific strategy. We started from a general specification including in all equations of the model the following variables, which are suggested by the previous literature and can be considered reasonably exogenous with respect to a child's education and smoking: smoking-awareness (i.e. knowledge that smoking can damage health), absence of mother, absence of father, mother's and father's interest in a child's education, mother's and father's smoking status, whether other members of the family smoke, child's ethnicity , mother's and father's education, British Ability Scales score (as a proxy of ability),[4] region of residence, home ownership, parents' weekly income bands in pounds sterling, highest social class between parents. The description of these vari-

[3]See the discussion in **?**) in the context of the multinomial probit model.

[4]See **?**). In the 10-year follow-up survey of the BCS70 not all individuals were administered the BAS test.

ables is included in Table 4 All covariates except HE, which is measured at age 29, are measured at age 10. In case variables had missing values, we included a missing value dummy.[5]

In this general model, which is formally identified, we ran exclusion Wald tests for various groups of variables in each equation.[6] Groups that were statistically significant at 10% or more for at least one gender were retained in the specifications for both genders, in order to obtain 'parsimonious' and comparable specifications across genders.

Through this variable selection process different covariates were chosen for inclusion in the HE and the smoking intensity equations. In particular, the final specification of the HE equation includes: absence of mother or father, mother's and father's interest in child's education, parental smoking status, other smokers in the family, parents' education, child's ethnicity, home ownership, BAS score. The final specification of the smoking intensity equation includes: the HE dummy, absence of mother or father, mother's and father's interest in child's education, parental smoking status, other smokers in the family, child's ethnicity, region of residence. Hence, the model is identified by more than one exclusion restriction. However, we consider parental education as the main 'identifying variable'. On the one hand, previous work suggests that the main influence of a parent's education on his/her children's smoking habits amounts to the transmission of health knowledge, differences in parenting styles and the role model transmitted through his/her own smoking habits. Hence, it seems reasonable to assume that, conditional on the smoking status of parents and proxies of parenting quality, parental educa-

[5]Exclusion of all observations with missing values for at least one covariate produces a large drop in the sample size. We checked the robustness of our results by estimating the model also in the samples with non-missing values and obtained qualitatively and quantitatively very similar results.

[6]To ease presentation these exclusion Wald tests are not reported here. The statistics are however available from the authors upon request. Following the suggestion of one anonymous referee and to avoid running the risk of omitting relevant variables, we retained parental interest in child's education in both the smoking intensity and the HE equations, although it was only significant in the latter. Indeed, this variable may be a proxy of parental interest in child's behaviour (i.e. parenting quality).

tion only affects children's smoking behaviour through its effect on children's education (see for example **???**). On the other hand, the correlation between parents' and children's education is well established in the intergenerational mobility literature (**?**). Following these arguments parental education was excluded from the smoking equation and included in the child's education equation.[7] Wald tests for the exclusion of parental education in the HE and smoking intensity equations that were peformed in the ED-OP model including the smaller set of covariates above supported this decision. For women the Wald test values were (distributed as a $\chi^2(10)$, p-values are in parentheses) 83.15 (0.00) and 13.40 (0.20) for the HE and the cigarette equation, respectively. For men the corresponding test values were 55.58 (0.00) and 8.95 (0.54), respectively.

Variables such as an individual's income or job qualifications were excluded from the smoking intensity equation since they are potentially endogenous. By excluding these variables we estimate the overall effect of HE on current smoking intensity conditional on current smoking participation and gross of the effect running from HE towards smoking through income and job qualification. However, in Section 3.3 we investigate whether the effect of education is mainly accounted for by job-related variables.

3.3 Main results

Each row of Table 5 reports the estimate obtained with a specific econometric model or including a specific set of covariates. Model (1) for both men and women reports the estimates of marginal effects (at the sample mean) for the ordered probit model with exogenous HE. The results are consistent across genders. That is, *among current smokers at age 29*, HE has a positive association with the probability of being an occasional smoker and a negative one with that of being a medium/heavy smoker. In particular, women (men) with HE are 13.2 (11.8) percent points (p.p.) more likely to be occasional

[7] The same identifying restriction is used, for instance, by **??**) to estimate the causal effect of education on smoking status.

smokers, 4.7 p.p. less likely to smoke between 11 and 15 cigarettes a day, 8.5 (8.2) p.p. less likely to smoke between 16 and 20 cigarettes, and 2.8 p.p. (5.3 p.p.) less likely to smoke more than 20 cigarettes.

[Table 5 about here]

Model (2) reports the estimates of the ED-OP model. Once the endogeneity of HE is accounted for, results show that the effect of HE on smoking is larger than that estimated from the ordered probit model with exogenous HE. Moreover, now statistically significant effects of HE on smoking also emerge at low levels of smoking intensity. The positive effect of HE on the probability of being an occasional smoker rises to 27.2 p.p. for women and 21.1 p.p. for men. HE also has positive effects on the probability of smoking 1-5 cigarettes for women (3.9 p.p.) and on the probability of smoking 1-5 cigarettes or 6-10 cigarettes for men (2.7 p.p. and 2.1 p.p., respectively). Also the effects at the top of the smoking distribution increase. Indeed, women with HE have a 9.2 , 15 and 4.9 p.p. lower probability of smoking 11-15, 16-20 and more than 20 cigarettes, respectively, than those with lower education. Similarly, men with HE turn out to be 3.9, 13.6 and 8.4 p.p. less likely to smoke 11-15, 16-20 and more than 20 cigarettes, respectively.

The ED-OP models reported show a significant positive correlation between unobservables entering the HE and smoking intensity equations i.e., a positive and significant ρ, for both males and females. This result may be driven by the fact that in our econometric analysis we are conditioning on current smoking participation, and we are neglecting potential individual self-selection into current smoking participation. Then, model (3) reports the marginal effects obtained using a three-equation model of current smoking participation, HE and smoking intensity, which we label the Selection Endogenous Dummy Ordered Probit (SED-OP) model. The marginal effects are very similar to the ones of model (2), and the model does not show any evidence of correlation between the unobservables affecting current smoking participation and those affecting HE.[8]

[8]The main feature of the SED-OP model is that an endogenous dummy enters both the

The results of a positive correlation between smoking intensity and HE and of a zero correlation between current smoking participation and HE are not obvious and only apparently contrasting with the negative correlation between smoking participation and HE unobservables generally posited by the economic theory (running for instance through the intertemporal discount rate).[9] We propose a very simple 'story' to explain this empirical puzzle. Let us assume that two kinds of unobservables enter the demand for HE and smoking (either participation or intensity), w_i and q_i, respectively:

$$w_i = h_i - r_i \tag{6}$$

$$q_i = h_i + r_i \tag{7}$$

where h_i is the individual's latent level of 'health stock' and r_i the unobservable level of the discount rate. We have assumed that the discount rate enters negatively the decision to invest in HE and positively the decision to smoke (as the literature does), while we have assumed that the health stock enters positively both decisions (i.e. healthier individuals have higher returns to education or lower costs of studying, and they trade-off the 'health stock' with health-related behaviour; that is, healthy individuals may afford to engage in unhealthy behaviour). In such a simplified framework $cov(w_i, q_i) = var(h_i) - var(r_i)$. Hence, in the overall population the sign of the correlation between the unobservables may be either positive or negative depending on the difference between the variances in individuals' health stocks and in the discount rates.

Let us first consider the 'ever smoke' decision, that is the decision to start to smoke, which usually takes place at very young ages and for which the relevant population is the whole population. Clearly, when we consider the

selection equation for smoking participation (probit) and the main equation for smoking intensity (ordered probit). The smoking participation equation includes all the covariates included in the smoking intensity equation. Technical details on this model are available in **?**).

[9] Although some more recent research shows that an individual's cognitive abilities, which tend to be positively correlated with higher levels of education, are positively related to higher risk-taking behaviour (**??**).

'ever smoke' decision it is likely that the variance in the health stock is low - most individuals were young and in good health when they had to make such decision - while there might be a substantial heterogeneity in the discount rates. Therefore, a negative correlation between the unobservables is likely to emerge. Consider now only individuals who started smoking and focus on their decision not to quit by a given age. Now, the relevant population is composed exclusively of those individuals who decided to enter into smoking activity at least once in their life. As a consequence, it is likely that in the subpopulation of 'starters' there is a much lower variance in the discount rates (since they all did start smoking) and a higher variance in their health status. Hence, a positive correlation among the unobservables could emerge in this case: individuals with a better health endowment may afford to smoke more and may also be the ones who achieved HE. Finally, consider the current smoking decision. The relevant population is as in the first case the whole population. However, unlike the two previous cases, now the population of current non-smokers is a mixture of the two very heterogeneous subpopulations of never-smokers (high discount rates) and quitters (low discount rates) while the population of current smokers is likely to be self-selected in terms of (relatively better) health endowment. This means that we would observe among current non-smokers both individuals with HE who never started smoking and individuals who started and quitted, due to their bad health, while we would observe among the current smokers only highly educated people with a relatively better health endowment, who can smoke more. As a result, the sign of the correlation between the unobservables affecting higher education and current smoking status becomes uncertain.

To explore these speculations we estimate gender-specific sequential probit models for: 1) HE and 'ever smoked'; 2) HE and 'not quit smoking'. For the smoking and the HE equations we use the same specification as in the ED-OP model. The results for the first model are reported in column (1) of Table 6 for women and men respectively. Curiously enough, while

the correlation between the unobservables affecting HE and 'ever smoked' is negative and significant, there is no causal effect of HE on the likelihood of ever started smoking. This is rather intuitive: since smoking initiation usually takes place at early ages, well before individuals complete HE, then it is very unlikely that HE has a causal effect on the 'ever smoke' decision. Hence, as far the latter is concerned, we find evidence consistent with the third variable hypothesis put forward in **?**) and **?**). When analysing an individual's decision whether to quit smoking or not, we also find some empirical evidence consistent with our story. Column (2) of Table 6 shows the effect of HE on the decision of not quitting smoking. HE has a negative effect on not quitting for both men and women (cf. **??**), while the correlation coefficient between the unobservables affecting not quitting and HE is positive - the same sign observed when the smoking intensity decision was analysed - and significant for women at 5% and positive and insignificant for men. A possible rationalisation for the causal effect is that highly educated individuals could be more sensitive to the deterioration of their health status and to health issues in general and therefore more likely to quit as they grow old. Moreover, also in this case, like in the case of smoking intensity, highly educated smokers who do not quit are probably those individuals who have a better health genetic endowment, a factor which positively affects both HE and smoking and that may drive the empirically observed positive correlation between the unobservables. Evidence consistent with this interpretation is provided, for instance, in **?**) who find that past improvements in health while smoking are positively correlated with current cigarette consumption. Similarly, **?**) find a negative correlation between the propensity of getting smoking-related diseases and the propensity of continuing to smoke. **?**) finds that smoking is a substitute for a newly discovered risk of cancer (i.e., a lower genetic endowment) measured in terms of cancer family history. These genetic traits may also positively affect the demand for education inducing a positive correlation among the unobservables affecting the two processes

(smoking and education).[10]

Last but not least, in order to check whether HE mainly exerts its effects through job-related variables, specification (4) in Table 5 includes, in the ED-OP model, labour market status (dummies for social class and dummies for being unemployed, in education or in a government training scheme, other out of the labour force or missing labour market status) and specification (5), in the same table, includes both labour market status and net monthly pay. Results show that health returns to HE do not seem to originate mainly from job-related channels. This finding is in line with the previous literature showing that occupation explains only a small part of the differences in individual health status or health-related behaviour by level of education (**??**).[11]

3.4 Higher education and smoking dynamics

We have shown until now that HE negatively affects current *smoking participation* and the *level of smoking intensity* at age 29. As a further robustness check, in this section we estimated a Selection Endogenous Dummy - Dynamic Ordered Probit (SED-DOP) model using the 2000 and the 2004 BCS70 follow-up surveys to investigate smoking dynamics between age 29 and age 34.[12] We selected only individuals who smoked in 2000 and estimate a four-equation model: one equation for smoking continuation between 2000 and 2004 (selection dummy), one equation for HE (endogenous dummy), one equation for smoking intensity in 2000 (initial conditions), and one equation for current smoking intensity in 2004 given the initial conditions. Hence,

[10] Other explanations are also possible. For instance, higher earning ability may increase the demand for both cigarettes and education via an income effect. However, in what follows we check for this possibility by including an individual's labour income in the smoking intensity equation.

[11] Following the suggestion of an anonymous referee, we also checked whether the effect of HE was mainly driven by the presence of young children within the household and did not find any evidence supporting this hypothesis.

[12] Considering a later follow-up survey also gives a more complete picture of the life-cycle health returns to higher education.

the latter ordered probit is dynamic because past smoking intensity enters the equation of current smoking intensity. Obviously, current smoking intensity is only observed if the selection dummy takes on the value of one. The endogenous dummy for HE enters all smoking (participation and intensity) equations and we let past smoking intensity enter the current smoking participation equation as well. The SED-DOP model allows for a non-zero correlation between the error terms entering all four equations and takes due care of the initial conditions problem in the dynamic equation for smoking intensity (for more on the initial conditions problem, see, **?**).[13]

The estimates are reported in Table 7. HE turns out to have a significant negative effect on current smoking intensity, i.e. it decreases the likelihood of observing the highest smoking intensity categories for both men and women, even after controlling for past smoking intensity. The effect of HE on smoking intensity appears to be similar over time (i.e. in 2000 and in 2004) and close in magnitude to that estimated in Table 5 with the static SED-OP model. Notice that, after controlling for past smoking intensity, HE has no statistically significant effect on smoking continuation. We estimated the same model omitting past smoking intensity in the smoking continuation equation and found HE to be, in this instance, highly statistically significant. Hence, in the dynamic model the effect of HE on continuous smoking participation is mediated by its influence on past levels of smoking intensity.[14] The correlation coefficients between the error terms show that the endogeneity of HE with respect to current or past smoking intensity cannot be generally rejected and that there is not a sample selection problem (the correlation between the errors of the current smoking intensity and the current smoking participation equations is zero), confirming the findings of the (static) SED-OP estimates.

[13] More details on this model can be found in **?**). The smoking participation equation, the smoking intensity and the HE equations include the same set of controls as the SED-OP model. The lagged smoking intensity equation includes the same controls as the current smoking intensity equation (see Section 3.2).

[14] This means that the negative effect of HE on smoking participation estimated in the static SED-OP model in Table 5 was mainly capturing that of past levels of smoking intensity.

4 Concluding remarks

The present paper studies the causal effect of higher education on smoking intensity. It addresses the potential endogeneity of higher education with respect to smoking intensity by estimating an Endogenous Dummy - Ordered Probit (ED-OP) model. The distinctive feature of this model is that an endogenous dummy for higher education (HE) enters the main ordered model for the number of cigarettes smoked.

Our estimates using the 29-year follow-up survey of the 1970 British Cohort Study show that higher education is endogenous with respect to cigarette consumption, that HE has a negative effect on smoking intensity and that job-related variables do not provide the main causal pathway for this effect. We find that these results are robust to controlling for the potential self-selection of individuals into current smoking participation, and to estimating a dynamic model in which current smoking levels depend on past smoking levels (an 'addiction model'). This last model also shows that the positive effect of HE on the decision to quit smoking is mainly mediated by lower past levels of smoking intensity (i.e., 'less addiction'), a result that would deserve further investigation.

Concluding, our findings show the existence of important 'health returns' to higher education in the UK. Such evidence may be important to inform the current debate on the funding of tertiary education. Indeed, although the recent increase in student fees and in the incidence of private funding of the HE system have been motivated by the high private economic returns to HE, our analysis shows important non-pecuniary returns to HE. This has two major implications. First, if these 'health returns', to which substantial savings in public health expenditures might be associated, are not considered by policy makers when deciding the level of funding of higher education, then the level of public support to higher education might be sub-optimal. Secondly, increasing higher education might turn out to be a very effective way of increasing public health by reducing people's engagement in health-damaging behaviour.

Tables and Figures

Table 1 Smoking habits in the 29-year follow-up survey of BCS70

Smoking habits	Frequency	Percent	Cumulate
never smoked cigarettes	4,937	44.06	44.1
used to smoke but don't at all now	2,125	18.97	63
smoke cigarettes occasionally	863	7.7	70.7
smoke cigarettes every day	3,279	29.27	100
Total	11,204	100	

Note. The distribution refers to valid answers only.

Table 2 Number of cigarettes smoked daily by education (current smokers)

Education at	Women			Men		
age 29	Mean	Std. Dev.	Freq.	Mean	Std. Dev.	Freq.
lower than HE	11.37	7.94	1350	13.45	9.45	1,529
HE	7.68	7.80	408	9.53	8.89	461
Total	10.51	8.06	1758	12.54	9.47	1,990

Note. Data refer to the 29-year follow-up survey of BCS70.

Table 3. Highest educational qualification of the lower than HE group

Composition	Women			Men		
	Freq.	Percent	Cum.	Freq.	Percent	Cum.
No formal qualification	336	24.96	24.96	350	23.01	23.01
Level 1: GSCE D-G, CSE 2-5,..	167	12.41	37.37	180	11.83	34.85
Level 2: O-level equiv.	639	47.47	84.84	667	43.85	78.7
Level 3: A-level equiv.	204	15.16	100	324	21.3	100
Total	1,346	100		1,521	100	

Note. Level 1 and Level 2 refer to compulsory schooling while Level 3 to post-compulsory schooling. Data refer to the 29-year follow-up survey of BCS70. O-levels are the (compulsory) lower secondary educational qualification in the UK, while A-levels education is entered at age 16, when individuals have completed compulsory schooling, typically by those individuals wishing to go on in HE and represent the upper secondary educational qualification.

Table 4. Variables description

Variable	Description	Categories[(a)]
Child's smoking-awareness	child's answer to: Can smoking damage your health?	may be true, yes I believe it, missing (I don't believe it)
Mother not present	dummy variable for mother not present	-
Father not present	dummy variable for father not present	-
Mother's interest in child education	mother's level of interest in child's education	little or no interest, cannot say, missing information (very interested)
Father's interest in child education	father's level of interest in child's education	little or no interest, cannot say, missing information (very interested)
Mother smokes	mother's smoking habit	non-smoker, missing (smoker)
Father smokes	father's smoking habit	non-smoker, missing (smoker)
Other smokers in the household	presence of other smokers within the household	no, missing (yes)
Ethnic group	child's ethnic group	non-European, missing (European)
Mother's education	mother's highest level of education	O-level, A-level, Professional, Degree, missing (less than O-level)
Father's education	father's highest level of education	O-level, A-level, Professional, Degree, missing (less than O-level)
BAS score	British Ability Scales score (verbal + quantitative)	-
BAS score missing	dummy variable for missing BAS score	-
Region	region of residence	North-East, North-West, Yorkshire and Humbershire, East Midlands, West Midlands, East, London, South West, Wales, Scotland (South East)
Home ownership	accommodation owned or rented	mortgage, rented or other, missing (owned)
Household income	combined gross parental weekly income bands (pounds)	35-50, 50-100, 150-200, 200-250, >250, missing (<35)
Social class	highest between parents' social classes	Intermediate, Skilled Non-manual, Skilled Manual, Partly Skilled, Unskilled, unemployed, out of the labour force (Professional)

Note. This table reports the definition for the control variables included in the econometric models. [(a)] Only for categorical variables. Omitted category in parentheses.

Table 5 Marginal effects of HE on smoking (intensity and participation) from different models and specifications

Models	*selection equation:* current smoker (P)	smoking intensity (S)					
			usual smoker				
		occasional smoker	no. of cigarettes per day				
			1–5	6–10	11–15	16–20	> 20
Women							
(1) Exogenous HE	-	0.132***	0.026	0.001	-0.047***	-0.085***	-0.028***
		(0.021)	(0.105)	(0.084)	(0.015)	(0.018)	(0.004)
(2) ED-OP	-	0.272***	0.039***	-0.019	-0.092***	-0.150***	-0.049***
		(0.058)	(0.004)	(0.013)	(0.018)	(0.023)	(0.010)
Corr(HE,S) – ρ					0.269***		
					(0.096)		
(3) SED-OP	-0.141***	0.317***	0.034***	-0.038**	-0.105***	-0.158***	-0.050***
	(0.034)	(0.053)	(0.005)	(0.017)	(0.015)	(0.019)	(0.009)
Corr(HE,S)			0.344*** (0.085)				
Corr(HE,P)			0.059 (0.065)				
Corr(S,P)			0.041 (0.056)				
(4) ED-OP with occupation	-	0.207***	0.036***	-0.007	-0.074***	-0.124***	-0.037***
		(0.059)	(0.006)	(0.011)	(0.020)	(0.027)	(0.009)
(5) ED-OP with occupation and wage	-	0.205***	0.036***	-0.007	-0.074***	-0.123***	-0.037***
		(0.059)	(0.006)	(0.011)	(0.020)	(0.027)	(0.009)
No. observations	5,188			1,754			
Men							
(1) Exogenous HE	-	0.118***	0.018	0.019	-0.020	-0.082***	-0.053***
		(0.019)	(0.100)	(0.085)	(0.024)	(0.013)	(0.007)
(2) ED-OP	-	0.211***	0.027***	0.021***	-0.039***	-0.136***	-0.084***
		(0.056)	(0.005)	(0.004)	(0.012)	(0.030)	(0.017)
Corr(HE,S) – ρ				0.183* (0.095)			
(3) SED-OP	-0.138***	0.218***	0.028***	0.023***	-0.039***	-0.140***	-0.089***
	(0.034)	(0.051)	(0.005)	(0.005)	(0.010)	(0.028)	(0.019)
Corr(HE,S)			0.207** (0.093)				
Corr(HE,P)			-0.032 (0.059)				
Corr(S,P)			-0.013 (0.054)				
(4) ED-OP with occupation	-	0.164***	0.023***	0.022***	-0.030**	-0.111***	-0.069***
		(0.057)	(0.006)	(0.003)	(0.013)	(0.033)	(0.019)
(5) ED-OP with occupation and wage	-	0.164***	0.023***	0.022***	-0.030**	-0.111***	-0.069***
		(0.059)	(0.006)	(0.003)	(0.013)	(0.034)	(0.020)
No. observations	4,954			1,980			

Note. Marginal effects at the sample mean obtained from different models and covariate specifications. Robust standard errors in parentheses. OP, ED-OP and SED-OP stand for Ordered Probit, Endogenous Dummy - Ordered Probit and Selection Endogenous Dummy - Ordered Probit, respectively. These models also include the other covariates listed in section 3.2. Corr(i,j) stands for the correlation between the error terms in equation i and in equation j.
***significant at 1%; **significant at 5%; *significant at 10%.

Table 6 Marginal effects on 'ever smoked' (ES) and 'not quit smoking' (NQS) — (sequential probit models)

	Ever smoked (1)	Not quit smoking (2)
Women		
HE	-0.012	-0.217***
	(0.044)	(0.058)
Corr(HE,ES)	-0.191**	-
	(0.072)	
Corr(HE,NQS)	-	0.152
		(0.098)
No obs.	4,954	2,882
Men		
HE	-0.005	-0.221***
	(0.042)	(0.060)
Corr(HE,ES)	-0.165**	-
	(0.068)	
Corr(HE,NQS)	-	0.220**
		(0.101)
No. observations	5,189	2,788

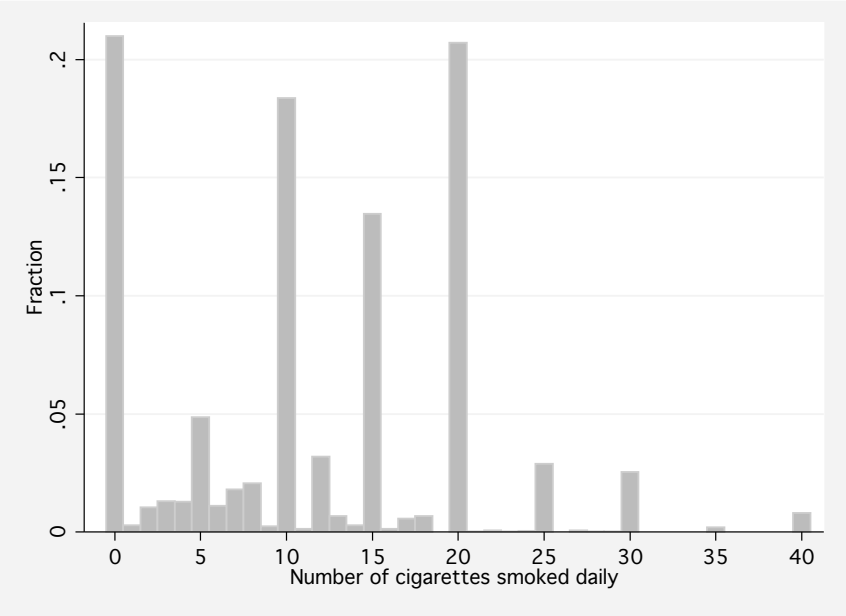
Note. Marginal effects at the sample mean obtained from sequential probit models. Robust standard errors in parentheses. These models also include the other covariates listed in section 3.2. Corr(i,j) stands for the correlation between the error terms in equation i and in equation j.
***significant at 1%; **significant at 5%.

Table 7. Marginal effects of HE on smoking from the SED-DOP model

	selection equation: smoking continuation (P_{2004})	smoking intensity (dynamic equation)					
			usual smoker				
		occasional smoker	no. of cigarettes per day				
			1–5	6–10	11–15	16–20	> 20
Women							
HE	0.058						
	(0.250)						
Current smoking intensity (S_{2004})							
HE		0.449***	0.050***	-0.030	-0.134***	-0.226***	-0.109***
		(0.048)	(0.009)	(0.02)	(0.013)	(0.019)	(0.022)
Lagged smoking intensity (S_{2000})							
HE		0.486***	0.029***	-0.067***	-0.140***	-0.209***	-0.099***
		(0.028)	(0.005)	(0.011)	(0.009)	(0.011)	(0.011)
Corr(HE,P_{2004})			-0.19 (0.04)				
Corr(HE,S_{2000})			0.622*** (0.022)				
Corr(HE,S_{2004})			0.613*** (0.048)				
Corr(S_{2000},P_{2004},)			-0.236 (0.622)				
Corr(S_{2004},P_{2004})			-0.232 (0.646)				
Corr(S_{2004},S_{2000})			0.762*** (0.273)				
No. observations	1,350			1,004			
Men							
HE	-0.041						
	(0.166)						
Current smoking intensity (S_{2004})							
HE		0.218*	0.059***	0.075***	-0.084	-0.220**	-0.049***
		(0.131)	(0.023)	(0.012)	(0.053)	(0.087)	(0.019)
Lagged smoking intensity (S_{2000})							
HE		0.393***	0.033***	0.010	-0.082***	-0.221***	-0.134***
		(0.038)	(0.004)	(0.010)	(0.009)	(0.013)	(0.016)
Corr(HE,C_{2004})			0.017 (0.283)				
Corr(HE,S_{2000})			0.542*** (0.044)				
Corr(HE,S_{2004})			0.308 (0.239)				
Corr(S_{2000},P_{2004})			0.019 (0.318)				
Corr(S_{2004},P_{2004},)			0.011 (0.241)				
Corr(S_{2004},S_{2000})			0.334 (0.213)				
No. observations	1,368			1,055			

Note. This model is estimated on those individuals who were current smokers in 2000. Marginal effects at the sample mean obtained using the Selection Endogenous Dummy - Dynamic Ordered Probit (SED-DOP) model. Robust standard errors in parentheses. These models also include the other covariates listed in section 3.2. Corr(i,j) stands for the correlation between the error terms in equation i and in equation j.
***significant at 1%; **significant at 5%; *significant at 10%.

Figure 1. Number of cigarettes smoked daily (current smokers)



Note. Data refer to current smokers in the 29-year follow-up survey of BCS70. Current smokers who do not smoke every day (i.e. occasional smokers) are attributed the value of zero.