HEDG
HEALTH, ECONOMETRICS AND DATA GROUP
THE UNIVERSITY of York

WP 15/07

Intelligence and the Mortality Difference by Education: Selection or mediation?

Govert E. Bijwaard & Andrew M. Jones

July 2015

http://www.york.ac.uk/economics/postgrad/herc/hedg/wps/

Intelligence and the Mortality Difference by Education: Selection or mediation?*

Govert E. Bijwaard† NIDI

Andrew M. Jones‡ University of York

July 4, 2015

Abstract

Large differences in mortality rates across those with different levels of education are a well-established fact. This association between mortality and education may partly be explained by confounding factors, including intelligence. Intelligence may also be affected by education so that it becomes a mediating factor in the causal chain. In this paper we estimate the impact of education on mortality using inverse probability weighted (IPW) estimator, using either intelligence as a selection variable or as a mediating variable. We develop an IPW estimator to analyse the mediating effect in the context of survival models. Our estimates are based on administrative data, on men born in 1944-1947 who were examined for military service in the Netherlands between 1961-1965, linked to national death records. For these men we distinguish four education levels and we make pairwise comparisons. From the empirical analyses we conclude that the mortality differences observed by education are only attributable to education effects for high educated individuals. For low educated individuals the observed mortality gain is mainly attributable to differences in intelligence.

JEL classification: C41, I14, I24.

Keywords: Education; Mortality; Inverse probability weighting; Mediators; Mixed proportional hazard

*The authors acknowledge access to linked data resources (DO 1995-2011) by Statistics Netherlands (CBS). We are grateful to seminar participants at Erasmus University Rotterdam, University of York and the Paris School of Economics for helpful comments.

†Netherlands Interdisciplinary Demographic Institute (NIDI-KNAW/University of Groningen), PO Box 11650, 2502 AR The Hague, the Netherlands and IZA, Bonn, +31 70 3565224, bijwaard@nidi.nl

‡Department of Economics and Related Studies, University of York, YO10 5DD, United Kingdom; Centre for Health Economics, Monash University, Australia; Department of Economics, University of Bergen, Norway, andrew.jones@york.ac.uk

1 Introduction

Many studies show large differences in health and mortality across educational groups. This is one of the most compelling and well established associations in social science research and holds across many populations. Even in an egalitarian country such as the Netherlands, with a very accessible health care system, the difference in life expectancy between individuals with no formal education beyond primary school and those with a university education is more than five years (Bruggink 2009). Still, the background of these inequalities is not fully understood.

The association between mortality and education may partly be explained by confounding factors such as intelligence and parental background that affect both education choices and mortality (Deary 2008). Lower intelligence as measured by standardized IQ tests is related to increased mortality (Batty and David 2004; Batty et al. 2007; Calvin et al. 2011). Because educational attainment and intelligence are strongly correlated, it is difficult to separate their effects on mortality (Deary and Johnson 2010).

Studies based on natural experiments in education, including changes in compulsory schooling laws, may to some extent overcome the difficulty of separating true education effects from these confounding factors. Recent analyses of such natural experiments suggest that the causal effect of education on health outcomes may be limited (Mazumder 2008; Jones et al. 2011; Van Kippersluis et al. 2011; Meghir et al. 2013; Basu et al. 2014) or even absent (Albouy and Lequien 2009; Clark and Royer 2013). This suggests that confounding factors may well play an important role in shaping the strong association between education and health.

Studies based on structural models in which the interdependence between education, health, and cognitive ability is explicitly modelled show that at least half of the health disparities across educational groups is due to selection of healthier, more able individuals into higher education (Conti and Heckman 2010; Conti et al. 2010; Bijwaard et al. 2015). In a structural multistate model, including the transitions into and out of hospital, Bijwaard and van Kippersluis (2015) show that the lower mortality hazard in hospital for the higher educated is mostly removed after accounting for differences in intelligence.

The effects of intelligence on mortality could be operating in several ways. Indirect effects can be expected if higher education improves intelligence which may lead to improvement in social economic status later in life. Direct effects are likely if more intelligent individuals do better in managing their diseases and in seeking appropriate treatment where necessary (Batty and David 2004). Education and intelligence may also operate in tandem and be mutually reinforcing. A better understanding of these relations is needed to establish potential direct benefits of improvements in education on

mortality. Failure to control for intelligence in health and mortality analyses biases the estimated effect of education. Further, the effect of intelligence on health and mortality is of direct interest as higher intelligence gives the higher educated their efficiency advantage in terms of health investment (Auld and Sidhu 2005; Bijwaard and van Kippersluis 2015).

In this paper we try to obtain the causal impact of improving education on the mortality risk while accounting for differences in intelligence. Ideally, we would have continuous measurement of the (development) of intelligence over the life cycle, to account for both the selection and mediation of intelligence in the causal path from education to mortality. However, we only observe intelligence at late adolescence when measured intelligence can be both the result of the attained education and a proxy of early childhood intelligence which influences education choice. We will investigate how these two opposing assumptions of the place of measured intelligence at age 18 in the causal path from education to mortality affects the estimated impact of education on mortality.

Our outcome, the age at death is a duration variable and the mortality hazard rate, the instantaneous probability that an individual dies at a certain age conditional on surviving up to that age, is modelled. Accounting for right-censoring, when the individual is only known to have survived up to the end of the observation window, and left-truncation, when only those individuals are observed who were alive at a certain time, are easy to handle in hazard models (Van den Berg 2001). A common way to accommodate the presence of observed characteristics is to specify a proportional hazard model, in which the hazard is the product of the baseline hazard, the age dependence, and a log-linear factor of included covariates. Neglecting confounding in inherently non-linear models such as proportional hazard models leads to biased inference.

Propensity score methods are increasingly used to estimate causal effects in observational studies, e.g. see Caliendo and Kopeinig (2008) for a survey. These methods aim to adjust for confounding factors between the treatment groups, i.e. different education levels. The advantage of the propensity score is that it enables us to summarize the many possible confounding covariates as a single score (Rosenbaum and Rubin 1983). Right censoring makes inference of differences in means, as is standard in treatment analysis, unreliable. Propensity score methods for hazard models have been introduced for duration data that account for censoring, truncation and dynamic selection issues (Cole and Hernán 2004; Austin 2014). We apply inverse probability weighting (IPW) using the propensity score (Hirano et al. 2003). IPW using the propensity score belongs to the larger class of marginal structural models that account for time-varying confounders when estimating the effect of time-varying covariates (Robins et al. 2000). Here we extend these methods to mediation analysis.

A main methodological contribution of this paper is that we disentangle the effect of education on

mortality into a direct and an indirect, running through intelligence, effect. We derive and implement an inverse probability weighting (IPW) estimator for estimating such direct and indirect effects in (M)PH models. The estimator identifies causal mechanisms given that a sequential unconfoundness condition holds. This implies that (i) the treatment (education choice) is exogenous given the measured confounders and (ii) the mediator (intelligence) is exogenous given the measured confounders and the treatment.

In our empirical analyses we use administrative data on Dutch men who were examined for military service in the Netherlands between 1961-1965 after completing their secondary schooling. We followed 45,037 men selected from the national birth cohorts 1944-1947. These examinations are based on yearly listings of all Dutch male citizens aged 18 years in the national population registers. The sampled examination records were linked by Statistics Netherlands to recent national death records (up to 2012). The records include a standardized recording of demographic and socioeconomic characteristics such as education, father's occupation, religion, family size, and birth order, along with a standardized psychometric test battery. Educational level was classified in four categories:primary school (age 6-12 years); lower vocational education (two years post primary school); lower secondary education (four years post primary school); and intermediate vocational education, general secondary education, higher non-university and university education (at least six years post primary school).

The empirical results show that improving education has hardly any impact on the mortality rate when accounting for intelligence. Only for the lowest education group we find a significant mortality reduction of 11% when these men would have improved their education. Using the mediation method we only find a significant indirect effect of education on mortality, running through intelligence, for this group that amounts to a 15% reduction in the mortality rate. For the highest education group we find a significant direct effect of education on mortality of 12%.

2 Data

Data from a large sample from the nationwide Dutch Military Service Conscription Register for the years 1961-1965 and male birth cohorts 1944-1947 are analysed. All men, except those living in psychiatric institutions or in nursing institutes for the blind or for the deaf-mute, were called to a military service induction exam. The majority attended the conscription examination at age 18. We have information from the military examinations for 45,037 men. The data were described elsewhere, (Ekamper et al. 2014), here we provide the main characteristics. These data were linked to the Dutch death register through the end of 2012 using unique personal identification numbers. Follow-up status was incomplete (due to emigration and other right-censoring events) for 1,316 (2.9%) and entirely

unknown for 2,625 (5.8%) men. The latter were removed from the data. These data allow us to follow a large group of men from age 18 till age 70 or till death. At the military examination a standardized recording of demographic and socioeconomic characteristics such as education, father's occupation, religion, family size, region of birth, and birth order is recorded. We exploit the information on education attained at age 18 and the age at death to investigate the mortality difference while accounting for other factors that both influence the education choice and the mortality.

Educational level was classified in four categories[1], (Doornbos and Kromhout 1990): primary school (age 6-12 years); lower vocational education (two years post primary school); lower secondary education (four years post primary school); and intermediate vocational education, general secondary education, higher non-university and university education (at least six years post primary school). For this study, we excluded partly institutionalized conscripts who had attended special schools for the illiterate, handicapped, deaf-mute, or mentally retarded, and conscripts who had not completed schooling through 12 years. After exclusion of these 2,614 conscripts, 39,798 men remain for analysis.

Also included is a standardized psychometric test battery comprising Raven Progressive Matrices, a nonverbal, untimed test that requires inductive reasoning about perceptual patterns, the Bennett Mechanical Comprehension test, and tests for Clerical Aptitude, Language Comprehension, Arithmetic and a Global comprehensive score, that combines all five tests. All tests were administered to over 95% of the population examined at induction. Scores for all tests were grouped in six levels from 1 (highest) to 6 (lowest). The test scores are highly correlated with Pearson's r values in the range of .63 to .76. Here, we only focus on the scores of the comprehensive test.

Selected demographic and socioeconomic characteristics at the time of military examinations by education level are given in Table 1. First born conscripts tend to have higher education. Father's occupation was classified into five categories: professional and managerial workers; clerical, self-employed and skilled workers; farmers; semi-skilled workers including operators, process workers and shop assistants; and labourers and miners. Fathers with unknown occupations were classified separately. Education level is also strongly related to father's occupation; men with the highest education tend to have fathers in professional or managerial occupations. The place of birth was categorized in four urbanization levels based on agrarian and total population size. This distinguishes rural communities (rural communities with 20% or more farming population), urbanized rural communities (rural communities with less than 20% farming population), towns (townships and cities with less than 100,000

[1]Education in the Netherlands is characterized by education years and by school level. There are two parallel streams in the educational system- general academic and vocational. Streaming choices are made at the end of primary school. Students in the vocational stream cannot directly enter university. Students with more than twelve years of education will nearly always be in the academic stream (Schröder and Ganzeboom 2014; Vrooman and Dronkers 1986).

inhabitants), and cities with populations of 100,000 or more. Men from rural areas are lower educated on average. The combined cognition measure is the Global comprehensive score. Not surprisingly, men with the highest education tend to do best on the comprehensive IQ test. Our principal measure of health is mortality with ages of death ranging from age 18 till age 70. The lowest education group has a 70% higher mortality.

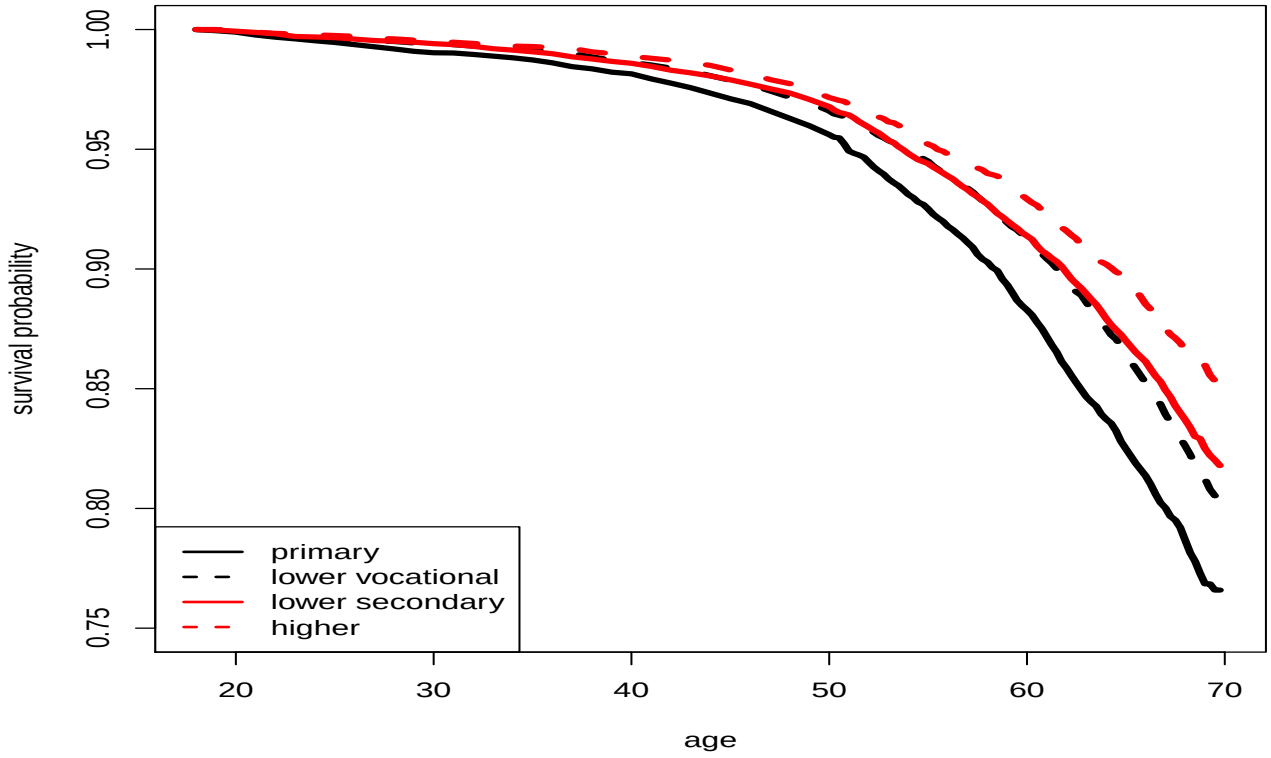
Table 1: Sample distribution by education level

	Primary education	Lower vocational	Lower secondary	Higher education	All levels
Birth order:					
1	27.8	32.1	39.3	42.6	35.5
2	27.1	30.3	30.7	29.9	29.9
3	18.7	18.4	16.3	15.4	17.3
4	11.3	9.2	6.9	7.0	8.4
≥ 5	14.9	10.0	6.7	5.1	8.8
Place of birth:					
City	76.0	74.4	82.1	83.3	78.6
Town	8.8	7.6	6.7	7.2	7.4
Urbanized Rural	2.8	2.7	2.0	1.7	2.3
Rural	12.5	15.3	9.2	7.8	11.7
Father's occupation:					
Professional	8.7	10.2	17.2	39.0	17.0
White collar	19.7	29.7	42.8	42.9	34.8
Farm owner	3.0	5.7	2.2	1.7	3.5
Skilled	38.4	33.3	23.1	9.2	26.7
Unskilled	22.5	14.9	9.4	3.4	12.3
Unknown	7.7	6.2	5.3	3.9	5.7
Global comprehensive IQ score:					
1 (highest)	0.1	6.3	19.8	54.6	17.6
2	3.8	27.5	47.9	37.7	32.5
3	13.7	30.3	20.9	4.0	20.6
4	28.3	22.7	7.2	0.6	14.9
5	39.5	10.6	1.7	0.1	10.1
6 (lowest)	11.5	0.8	0.1	0.02	2.0
Total # of deaths	1,213	2,522	2,109	827	5,350
% died	21.2	17.3	16.1	12.9	16.8
Sample size	5,712	14,572	13,124	6,390	39,798

The Kaplan-Meier survival curves for the four education categories primary lower vocational, higher vocational, and higher education, shown in Figure 1, reflect these mortality differences. Survival increases with the education level and the differences between the education levels increase with age. The curves differ significantly ($\chi^2 = 147.61$ for a log-rank test with 3 degrees of freedom). In subgroup analyses, survival differences comparing adjacent education levels are also statistically significant ($\chi^2 = 45.77, 5.79, 28.72$). This mortality difference by education is not necessarily due to

education per se. It could be that the higher intelligence of high educated people causes the difference. For example, understanding doctor's advice and adhering to complex treatments may be driven by intelligence rather than education. From Table 1 we have seen already that education and IQ are highly correlated. Figure 2 shows that survival also increases with IQ and the differences are statistically significant ($\chi^2 = 239.54$ for a log-rank test with 5 degrees of freedom). For all, except the two lowest, adjacent IQ levels the differences in the Kaplan-Meier survival curves are significant. Within each education level the Kaplan-Meier curves also differ significantly by IQ-level (not shown).

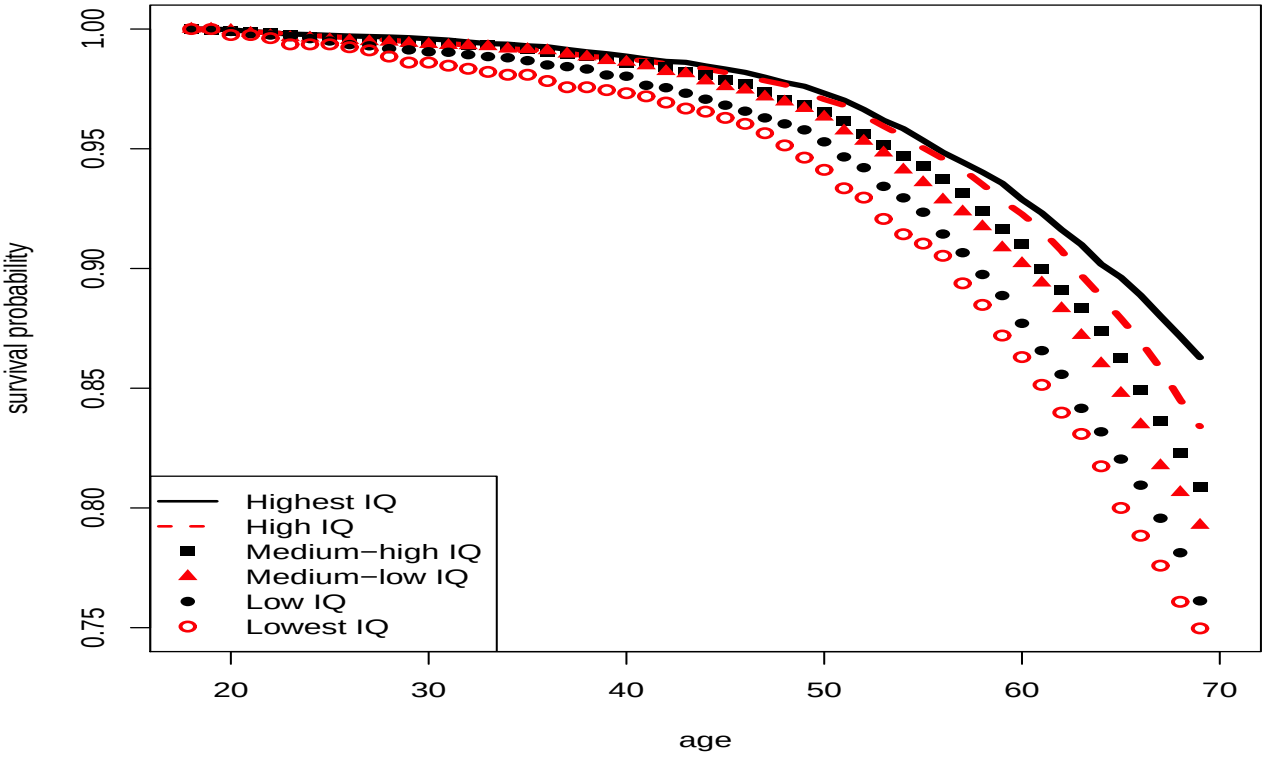
Figure 1: Kaplan-Meier survival curves, by education level



Next we investigate the relationship between IQ and attained education. The IQ scores are measured on a six-point ordinal scale. Comparing individuals on the extremes of the education level is not helpful as these individuals differ too much in many respects. We focus on adjacent education levels only and estimate separate ordered probit models for the IQ-score in relation to the highest education level in each pair and other observed individual characteristics. The results of ordered probit analyses reveal a strong association between education and IQ.[2]

[2]See Table B.1 in Appendix B.

Figure 2: Kaplan-Meier survival curves, by IQ level (overall level)



3 Defining the effect of education on the mortality hazard rate

We seek to find the impact of education level on the mortality risk for the men in our sample of conscripts. However, mortality may be influenced by factors that also determine the education choice. This may render education a selective choice and makes it endogenous to the mortality later in life. We follow a propensity score method to account for selection on observed characteristics and estimate the effect of education on the mortality rate. From the descriptive analyses in the previous section is it obvious that intelligence, measured by an IQ-test, influences both the education attained and the mortality later in life. However, intelligence as measured at age 18, the age at military examination, is also influenced by the education taken up till that age (Ceci 1991; Hansen et al. 2004; Carlsson et al. 2015). Figure 3 provides a graphical illustration of the relationship between intelligence, education and mortality later in life using a directed acyclic graph, where each arrow represents a causal path (Pearl 2000; Pearl 2012). It states that early childhood characteristics X, such as parental background and family size, influence the education choice D, the early childhood intelligence, IQ_0, and the intelligence at age 18, IQ_{18}. The latter is also influenced by early life intelligence and the education followed up till age 18. Unfortunately, we do not observe the early childhood intelligence.

We will investigate how different assumptions of the place of intelligence at age 18 in the causal path from education to mortality affects the estimated impact of education on mortality. The most

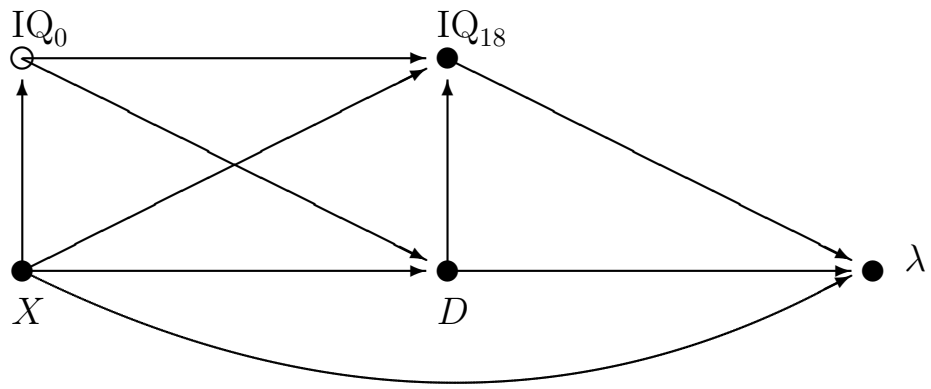


Figure 3: Directed acyclic graph of possible relation between IQ, education and mortality

simple model ignores intelligence in both the education choice and the mortality. In Section 4 we show that such an analysis will overestimate the impact of education on mortality. Next, in Section 5, we assume that intelligence at age 18 is a proxy for the intelligence early in life and is one of the factors that influence both the education choice and the mortality and show that this renders the impact of education insignificant. An alternative is to assume that education raises intelligence and that part of the impact of education on mortality runs through increased intelligence. In Section 6 a model in which intelligence at age 18 mediates the impact of education on mortality is introduced. This allows us to identify the direct and indirect effect of education on mortality running through intelligence. Before we discuss these models we define how we measure the impact of education on mortality.

We define the treatment effect, of moving up one education level, in terms of a proportional change in the (mortality) hazard rate. First, we discuss the assumptions, common in the potential outcomes literature using propensity score methods, to identify the impact of education on mortality risk. In Section 6 we extend this to identify the direct effect of education on the mortality rate when intelligence is treated as a mediator of this effect. The main difference with standard propensity score methods is that we use potential hazard rates, the hazard rate that would be observed if the individual was untreated, $\lambda(t|0)$, or treated $\lambda(t|1)$. Let $D_i = 1$ be the treatment, moving up one education level. We observe pre-treatment (education) covariates X that influence the education choice.

Assumption 1. *Unconfoundedness:* $\lambda(t|d) \perp D|X$ for $d = 0, 1$

where $\perp$ denotes independence. The unconfoundedness assumption (Rubin 1974, Rosenbaum and

Rubin 1983) asserts that, conditional on covariates X, treatment assignment (education level) is independent of the potential outcomes. This assumption requires that all variables that affect both the mortality and the education choice are observed. Note that this does not imply that we assume all relevant covariates are observed. Any missing factor is allowed to influence either the outcome or the education choice, not both. Although this is not testable and clearly a strong assumption, it may be a reasonable approximation. Any alternative, that does not rely on unconfoundedness while allowing for consistent estimation of the average treatment effects, will have to make alternative untestable assumptions. We check the robustness of our estimates to this unconfoundness assumption by assessing to what extent the estimates are robust to violations from this assumption induced by an additional binary variable, in Section 4.1 for the simple model ignoring intelligence, in Section 5.1 for the model including intelligence and in Section 6.2 for the (extended) model treating intelligence as a mediator.

Assumption 2. *Overlap:* $0 < \Pr(D = 1|X) < 1$.

The overlap, or common support assumption requires that the propensity score, the conditional probability to choose a higher education given covariates X is bounded away from zero and one. This assumption is in principle testable. If there are values of the covariates for which the probability of choosing a higher education level is zero or one, we cannot compare the 'treated' and 'control' individuals at these values. In that case we have to limit comparisons to sets of values where there is sufficient control in the propensity score among treated and controls. In our data we distinguish four (ordered) education levels in line with the contemporary Dutch education system (see Section 2). By comparing only adjacent education levels we remove most of the overlap problems.

We are also interested in estimating the average treatment on the treated (ATT) and the average treatment on the untreated (ATU). The ATT provides the average effect of education on mortality to those who obtained a higher education level, while the ATU provides the average effect of education for the lower educated had they obtained a higher education level. We can weaken the two assumptions in both instances. When interested in the ATT:

Assumption $1'$**.** *Unconfoundedness for controls:* $\lambda(t|0) \perp D|X$

Assumption $2'$**.** *ATT Overlap:* $\Pr(D = 1|X) < 1$.

While, if we are interested in estimating the ATU, the two assumptions can be weakened to:

Assumption $1''$**.** *Unconfoundedness for treated:* $\lambda(t|1) \perp D|X$

Assumption $2''$**.** *ATU Overlap:* $\Pr(D = 1|X) > 0$.

Rosenbaum and Rubin (1983) show that if the potential outcomes are independent of treatment conditional on covariates X, they are also independent of treatment conditional on the propensity score, $p(x) = \Pr(D = 1|X = x)$. Hence if unconfoundedness holds, all biases due to observable covariates can be removed by conditioning on the propensity score (Imbens 2004). The average effects can be estimated by matching or weighting on the propensity score. Here we use weighting on the propensity score. Inverse probability weighting based on the propensity score creates a pseudo-population in which the education choice is independent of the measured confounders. The pseudo-population is the result of assigning to each individual a weight that is proportional to the inverse of their propensity score. The inverse probability weighting (IPW) estimation is usually based on normalized weights that add to unity. Suppose we have a sample of n individuals, then based on an estimation of the propensity score, $\hat{p}(x)$, an estimator of ATE, ATT and ATU are all of the form $\sum_i [W_i \cdot D_i \cdot Y_i - W_i \cdot (1 - D_i) \cdot Y_i]$ with weights:

$$W_{i,\text{ATE}} = \left[\frac{D_i}{\hat{p}(X_i)} \Big/ \sum_{j=1}^{n} \frac{D_j}{\hat{p}(X_j)}\right] + \left[\frac{(1 - D_i)}{1 - \hat{p}(X_i)} \Big/ \sum_{j=1}^{n} \frac{1 - D_j}{1 - \hat{p}(X_j)}\right] \tag{1}$$

$$W_{i,\text{ATT}} = D_i + \left[\frac{(1 - D_i) \cdot \hat{p}(X_i)}{1 - \hat{p}(X_i)} \Big/ \sum_{j=1}^{n} \frac{(1 - D_j) \cdot \hat{p}(X_j)}{1 - \hat{p}(X_j)}\right] \tag{2}$$

$$W_{i,\text{ATU}} = \left[\frac{D_i \cdot \big(1 - \hat{p}(X_i)\big)}{\hat{p}(X_i)} \Big/ \sum_{j=1}^{n} \frac{D_j \cdot \big(1 - \hat{p}(X_j)\big)}{\hat{p}(X_j)}\right] + (1 - D_i) \tag{3}$$

The intuitive explanation of these weights can be best understood through the concept of pseudo-population. For example, each man with the highest education, $D = 1$, has a probability of $\hat{p}(X)$ to attain this education level. Similarly, a man with the lowest education, $D = 0$, has a probability of $1 - \hat{p}(X)$ to attain this education level. Therefore, in terms of the outcome distribution the outcome for those with $D = 1$ weighted by $1/\hat{p}(X)$ and for those with $D = 0$ weighted by $1/\big(1 - \hat{p}(X)\big)$ represents the original population from which the man is sampled. A problem with these basic weights is that they do not necessarily add up to one. We therefore normalise the weights to unity.

In survival analysis it is standard to compare the (non-parametric) Kaplan-Meier curves for the treated and the controls. The unadjusted survival curves may be misleading due to confounding. Cole and Hernán (2004) describe a method to estimate the IPW adjusted survival curves. Biostatisticians usually focus on Cox regression models and Cole and Hernán (2004) describe how Cox proportional

hazard models can be weighted by the inverse propensity score to estimate causal effects of treatments. This method is related to the g-computation algorithm of Robins and Rotnitzky (1992) and Robins et al. (2000). The standard Cox model, without additional covariates, assumes that the hazard is:

$$\lambda(t|D) = \lambda_0(t)\exp(\gamma D) \tag{4}$$

where $\lambda_0(t)$, the duration or age dependence, is left unspecified. The partial likelihood method (see e.g. Kalbfleisch and Prentice (2002)) provides an estimate of γ. The IPW Cox model is based on the weighted Cox partial likelihood score for γ:

$$U(\gamma) = \sum_{i=1}^{N} \delta_i W_i \left[D_i - \frac{\sum_j Y_j(t_i) W_j D_j exp(\gamma D_j)}{\sum_j Y_j(t_i) W_j exp(\gamma D_j)} \right] \tag{5}$$

where $Y_j(t_i) = I(t_j \geq t_i)$, the indicator that individual j is in the risk set at time t_i, the 'standard' counting process at-risk indicator (see Appendix A and the references therein) and δ indicates whether the duration for individual i is censored $\delta_i = 0$ or not. The IPW estimator of γ solves $U(\gamma) = 0$ and is proven to be a consistent if the model for the propensity score is correctly specified and the Cox model holds (Robins 1999).[3] Note that in a proportional hazard context it is natural to define the treatment effect proportionally, i.e. $e^{\gamma} = \lambda(t|D = D^1)/\lambda(t|D = D^0)$ instead of as a difference.

In economics the interest is often also in the duration dependence. The Gompertz hazard, which assumes that the hazard increases exponentially with age, $\lambda_0(t) = e^{\alpha_0 + \alpha_1 t}$, is known to provide accurate mortality hazards (Gavrilov and Gavrilova 1991). However, it is hardly ever possible to include all relevant factors, either because the researcher does not know all the relevant factors or because it is not possible to measure then. Ignoring such unobserved heterogeneity or frailty may have a huge impact on inference in proportional hazard models, see e.g. Van den Berg (2001). A common solution is to use a Mixed Proportional Hazard (MPH) model. it is assumed that all unmeasured factors and measurement error can be captured in a multiplicative random term V. The hazard rate becomes[4]

$$\lambda(t|D, V) = V\lambda_0(t)\exp(\gamma D), \tag{6}$$

The (random) frailty $V > 0$ is time-independent and independent of the observed characteristics X and treatment D. Note that independence of V and D is crucial, otherwise Assumption 1 would be violated. So, we assume that some factors influencing the mortality rate are not observed and that these factors do not influence the education choice. In the empirical application it is assumed that V has a gamma-distribution; a common assumption used in the empirical literature.

[3] See Appendix A for an alternative proof.

[4] A Cox MPH model is also possible, but harder to estimate. We focus on a Gompertz MPH model.

To adjust for confounding we estimate a standard MPH model, that does not include the measured confounders as covariates, using the re-weighted pseudo-population. Fitting a (mixed) proportional hazard model in the pseudo-population is equivalent to fitting a weighted MPH model in the original sample. The parameters of such weighted MPH models can be used to estimate the causal effects of education on mortality in the original sample. The IPW estimator in the (M)PH model is equivalent to solving the weighted derivatives of the log-likelihood:

$$U(\theta) = \sum_{i=1}^{N} W_i \left[\delta_i \frac{\partial \log \lambda(t_i|\cdot)}{\partial \theta} - \frac{\partial \Lambda(t_i|\cdot)}{\partial \theta} \right] \tag{7}$$

where θ is the vector of parameters of the hazard in (6) and $\Lambda(t|\cdot) = \int_0^t \lambda(s|\cdot)\, ds$, the integrated hazard.[5]

4 Model ignoring the influence of intelligence

First we ignore that intelligence may influence both the education choice and the mortality later in life. The framework presented so far considers only two possible states for each individual, either treated or untreated. This is too restrictive for our application. In principle we could estimate an ordered probit propensity score for our four ordered educational choices, see Imai and van Dyk (2004) and Feng et al. (2012). However, men in the lowest and highest education group differ too much in their observed covariates and IQ (see Section 2), which causes severe overlap problems (contradicting Assumption 2). We therefore define the effect of education through pairwise comparisons (Lechner 2002) of adjacent education levels: primary to lower vocational, lower vocational to lower secondary and lower secondary to higher education. For each comparison we estimate a separate propensity score of attaining the highest education level, (see Table B.2 in Appendix B). We included variables that influence both the propensity score and mortality. The included variables in the propensity score are the father's occupation, family size, regional dummies, famine birth cohorts and health indicators. For all three education comparisons the occupation of the father plays a crucial role in the propensity score. Religion only influences the education choice of the lowest educated. Health indicators at the military examination are also related to the education attained.

We estimate a (unadjusted and weighted) Cox model, a Gompertz model and Gompertz model with Gamma distributed unobserved heterogeneity (Gamma-Gompertz model). Table 2 presents the estimated effect on the mortality hazard of moving up one educational level for the Cox and the Gamma-Gompertz model (the results for the PH Gompertz model were very close to the results of the PH Cox model and are therefore not shown here). We conclude from these analyses that for the

[5] In Appendix A we provide a counting process interpretation and prove consistency.

lower educated, with only primary education, and for the lower secondary educated obtaining more education clearly reduces their mortality rate (around 20%). Moving from lower vocational education to lower secondary education does not change the mortality rate (except when looking at the treatment effect on the treated). Not adjusting for selective education choice seems to overestimate the impact of education. The treatment effect on the untreated is larger than the treatment effect on the treated. Thus inducing the lower educated to get more education would lead to a higher gain. Accounting for unobserved heterogeneity increases the impact of education for the men moving from lower secondary to higher education, see second panel of Table 2.

Table 2: Impact of education levels on the mortality rate

		IPW estimate		
	Unadjusted	ATE	ATT	ATU
Cox				
Primary to	-0.236^{**}	-0.185^{**}	-0.160^{**}	-0.218^{**}
lower vocational	(0.035)	(0.039)	(0.040)	(0.038)
Lower vocational to	-0.071^{+}	-0.052	-0.070^{+}	-0.046
lower secondary	(0.030)	(0.031)	(0.032)	(0.032)
Lower secondary to	-0.220^{**}	-0.190^{**}	-0.169^{**}	-0.197^{**}
higher	(0.041)	(0.046)	(0.044)	(0.049)
Gamma-Gompertz				
Primary to	-0.245^{**}	-0.185^{**}	-0.160^{**}	-0.230^{**}
lower vocational	(0.042)	(0.039)	(0.040)	(0.045)
Lower vocational to	-0.071^{+}	-0.052	-0.070^{+}	-0.046
lower secondary	(0.030)	(0.031)	(0.032)	(0.032)
Lower secondary to	-0.231^{**}	-0.216^{**}	-0.190^{**}	-0.221^{**}
higher	(0.045)	(0.053)	(0.051)	(0.055)

$^{+}p < 0.05$ and $^{**}p < 0.01$

For an IPW method to hold we need to check if the propensity score is able to balance the distribution of all included variables in both the control and treated group. One suitable way to check whether there are still differences is by calculating the standardized bias, or normalised difference in means:

$$100 \cdot \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{0.5\Big(\mathrm{Var}(x_1) + \mathrm{Var}(x_0)\Big)}} \tag{8}$$

Table 3 shows the percentage bias measure before and after adjusting the data in our sample. They reveal substantial imbalances between those who attained adjacent education levels, with the percentage bias as high as 48.8 percent for a father with a professional occupation in the highest education comparison. The biases in columns labelled 'after' show that these imbalances disappear for the non-IQ characteristics when we use the inverse propensity weights. However, the outcomes of the IQ-test are not balanced after weighting with propensity scores that ignore this test. This indicates we should include the IQ-test in the propensity score.

Table 3: Standardized bias before and after matching, pairwise comparisons

	Primary to lower vocational		Lower vocational to lower secondary		Lower secondary to higher	
	Before	After	Before	After	Before	After
Father's occupation						
Professional	4.9	−0.6	20.2	−0.2	48.8	−0.3
self-employed	−0.9	−0.3	5.4	−0.0	−10.1	−0.3
Skilled	−9.8	0.2	−22.9	0.0	−38.9	0.5
Unskilled	−20.4	0.4	−17.0	−0.1	−24.2	0.1
Missing	−5.9	−0.7	−4.8	0.1	−6.1	0.1
Family size	7.6	−0.2	7.5	−0.1	−2.4	0.7
Born in Utrecht	−11.9	0.0	2.4	−0.1	3.8	−0.4
Religion						
Catholic	−16.1	−0.5	−4.9	0.2	2.3	0.5
Dutch Reformed	12.6	0.8	0.7	−0.1	−2.5	0.0
Calvinist	17.4	−1.1	3.5	0.1	2.2	0.6
Other religion	−0.3	−0.1	3.1	−0.2	2.4	−0.9
Health						
Bad general health	−9.2	−0.7	6.9	0.1	−1.0	−0.1
Bad hearing	−8.6	0.4	−5.1	0.1	−2.2	−0.3
Bad sight	9.9	−0.2	29.6	−0.2	21.6	−0.8
Bad psychological	−35.9	−0.2	−2.2	0.1	−3.9	0.1
Famine cohorts:						
A1	4.6	0.0	0.0	0.0	−0.8	−0.2
A2	2.7	−0.3	−0.3	−0.1	−0.8	−0.4
B1	2.8	−0.4	2.7	0.0	0.5	−0.5
B2	4.1	0.0	2.4	−0.2	−3.6	0.6
D1	−2.8	−0.3	0.4	0.0	−2.9	−0.3
D2	0.6	−0.2	3.7	0.1	4.8	−0.1
Comprehensive IQ:						
1 (highest)	−35.0	34.8	40.8	34.3	77.1	73.3
2	69.3	65.4	43.0	38.6	−20.9	−19.4
3	41.0	36.1	−21.7	−18.8	−20.9	−51.3
4	−12.9	−16.6	−44.4	−40.1	−34.5	−33.6
5	−70.8	−67.6	−37.9	−34.3	−16.6	−15.2
6 (lowest)	−45.5	−37.0	−10.6	−9.2	−3.5	−2.6
# obs	20,272		27,687		19,497	

4.1 Sensitivity analyses

The critical assumption in propensity score weighting is that of no selection on unobservables. To test the sensitivity of matching estimators to the unconfoundness assumption we build on the sensitivity analyses of Nannicini (2007) and Ichino et al. (2008), adjusted for the use of a Cox-model.[6] The sensitivity analyses of Ichino et al. (2008) assume that the possible unobserved confounding factors can be summarised in a binary variable, U, and that the unconfoundness assumption holds conditional on X and U, i.e. $\lambda(t|0) \perp D|X, U$. Given the values of the probabilities that characterize the distribution of U we can simulate a value of the unobserved confounding factor for each individual and re-estimate the IPW-Cox. The probabilities of the distribution of U depend on the value of the treatment and the outcome. The Ichino et al. (2008) sensitivity analyses assumes that the potential outcomes are binary, but Nannicini (2007) shows how to extend this to continuous outcomes by imposing a binary transformation. In survival analysis we have a natural binary transformation, the censoring indicator $\delta_i = 1$ if individual i is still alive at the end of the observation period. Then, the distribution of the unobserved binary confounding factor U can be characterised by specifying the probabilities in each of the four groups.

$$p_{ij} = \Pr(U = 1|D = i, \delta = j, X) = \Pr(U = 1|D = i, \delta = j) \tag{9}$$

for $i, j = 0, 1$.

A measure of how the different configurations of p_{ij} chosen to simulate U translate into associations of U and the outcome and the treatment is ω, with ω the coefficient of U in a Cox model for the control group ($D = 0$) using U and X as covariates. Ichino et al. (2008) call this (exponentiated) the 'outcome effect'. A measure of the effect of U on the relative probability to be assigned to the treatment is ξ, with ξ the coefficient of U in a logit model on the treatment assignment ($D = 1$) using U and X as covariates. Ichino et al. (2008) call this (exponentiated) the 'selection effect'. Next we re-estimate the IPW-Cox treatment effects including U in the propensity score. The probability values of the distribution we impose on U are chosen such that they mimic the distribution for each included binary variable. For example, for the lowest education group (primary and lower vocational education) the probability that the individual is catholic is 0.413 for those with primary education and who died before the end of the observation period, p_{00}, 0.401 for those with primary education and who survived till the end, p_{01}, 0.319 those with lower vocational education and who died before the end, p_{10}, and 0.309 for those with lower vocational education and who survived till the end p_{11}. For each probability configuration of U we repeat the simulation of U the estimation of the outcome effect, selection effect

[6] Here we only focus on the effect in the Cox model. The methods can easily be extended to the Gompertz model or the Gompertz model with unobserved heterogeneity.

and the IPW-Cox treatment effects $M = 100$ times and obtain the average of these 100 simulations. The total variance of these averages can be estimated from (see Ichino et al. (2008))

$$\text{Var}_f = \frac{1}{M}\sum_{m=1}^{M} s_m^2 + \frac{M-1}{M(M-1)}\sum_{m=1}^{M}(\hat{f}_m - \bar{f})^2 \tag{10}$$

with $f \in \{\omega, \xi, \text{ATE}, \text{ATT}, \text{ATU}\}$ of each pairwise education comparison, $\hat{f}_m$ is the estimated f in simulation sample m and s_m^2 is its estimated variance.

Table 4: Sensitivity analysis IPW estimation of effect of education: outcome and selection effects

	Primary to lower vocational		Lower vocational to lower secondary		Lower secondary to higher	
	ω	ξ	ω	ξ	ω	ξ
1 (highest)	-4.45	3.94^{**}	-0.44^{**}	1.30^{**}	-0.16	1.58^{**}
	(12.41)	(0.55)	(0.15)	(0.06)	(0.09)	(0.05)
2	-0.43	2.28^{**}	-0.14^{+}	0.89^{**}	-0.08	-0.42^{**}
	(0.29)	(0.12)	(0.07)	(0.04)	(0.07)	(0.05)
4	-0.05	-0.29^{**}	0.08	-1.32^{**}	0.28^{+}	-2.55^{**}
	(0.10)	(0.06)	(0.07)	(0.06)	(0.11)	(0.25)
5	0.12	-1.71^{**}	0.20^{+}	-1.95^{**}	0.40	-2.84^{**}
	(0.13)	(0.06)	(0.09)	(0.11)	(0.22)	(0.66)
6 (lowest)	0.15	-2.75^{**}	0.20	-2.23^{**}	-0.52	-0.81
	(0.23)	(0.16)	(0.32)	(0.48)	(7.19)	(1.09)

Based on adding U to propensity score with probabilities of U from observed probabilities for comprehensive IQ. No effect would give $\omega = 0$ and $\xi = 0$. $^{+}p < 0.05$ and $^{**}p < 0.01$

Table 5: Sensitivity of impact of education

	Primary to lower vocational			Lower vocational to lower secondary			Lower secondary to higher		
	ATE	ATT	ATU	ATE	ATT	ATU	ATE	ATT	ATU
1 (highest)	-0.191	-0.177	-0.195^{**}	-0.017	-0.019	-0.026	-0.123^{+}	-0.116^{+}	-0.124^{+}
	(0.123)	(0.163)	(0.038)	(0.033)	(0.037)	(0.034)	(0.056)	(0.057)	(0.066)
2	-0.106	-0.060	-0.187^{**}	-0.030	-0.043	-0.029	-0.181^{**}	-0.177^{**}	-0.179^{**}
	(0.064)	(0.084)	(0.041)	(0.033)	(0.036)	(0.035)	(0.047)	(0.045)	(0.051)
4	-0.186^{**}	-0.163^{**}	-0.214^{**}	-0.022	-0.058	0.001	-0.186^{+}	-0.148^{**}	-0.201^{+}
	(0.039)	(0.041)	(0.038)	(0.034)	(0.034)	(0.039)	(0.068)	(0.046)	(0.088)
5	-0.142^{**}	-0.123^{+}	-0.159^{**}	-0.021	-0.051	-0.004	-0.187^{**}	-0.161^{**}	-0.196^{**}
	(0.046)	(0.052)	(0.049)	(0.036)	(0.034)	(0.044)	(0.058)	(0.045)	(0.071)
6 (lowest)	-0.165^{**}	-0.143^{**}	-0.192^{**}	-0.047	-0.068^{+}	-0.038	-0.190^{**}	-0.168^{**}	-0.197^{**}
	(0.043)	(0.044)	(0.060)	(0.032)	(0.032)	(0.035)	(0.046)	(0.044)	(0.049)

Based on adding U to propensity score with probabilities of U from observed for comprehensive IQ-test. Original estimates are ATE: -0.185 (Primary to lower voc.); -0.052 (Lower voc. to lower sec.); -0.190 (Lower sec. to higher); ATT: -0.160 (Primary to lower voc.); -0.070 (Lower voc. to lower sec.); -0.169 (Lower sec. to higher); ATU: -0.218 (Primary to lower voc.); -0.046 (Lower voc. to lower sec.); -0.197 (Lower sec. to higher).

Table 4 gives the predicted outcome and selection effects when U follows the distribution of each of the values of the IQ-test and Table 5 gives how the IPW estimates change when including these U

into the propensity scores.[7]. The estimated outcome effect ranges from -4.5 to 0.4 and selection effects range from -2.8 to 3.9. Only a few outcome effects are significant, but most of the selection effects. The implied IPW estimations using an IPW Cox model including the additional variable U reveal that the impact of education on mortality of may change with more than eight percentage point with respect to the baseline (not including U). Some of the estimated impacts have lost their significance. The largest change occurs for the IQ extremes. This seems to indicate that the results using a model ignoring the impact of intelligence are not robust and most overestimate the effect of education on mortality.

5 Model with intelligence included in the selection factors

From the previous analysis it is clear the IQ at age 18 is a source of selection into education. Still, we do not observe the intelligence at early age. Next, we assume that intelligence at age 18 is a proxy for the intelligence early in life and is one of the factors that influence both the education choice and the mortality. The selection model we assume is illustrated by the DAG in Figure 4

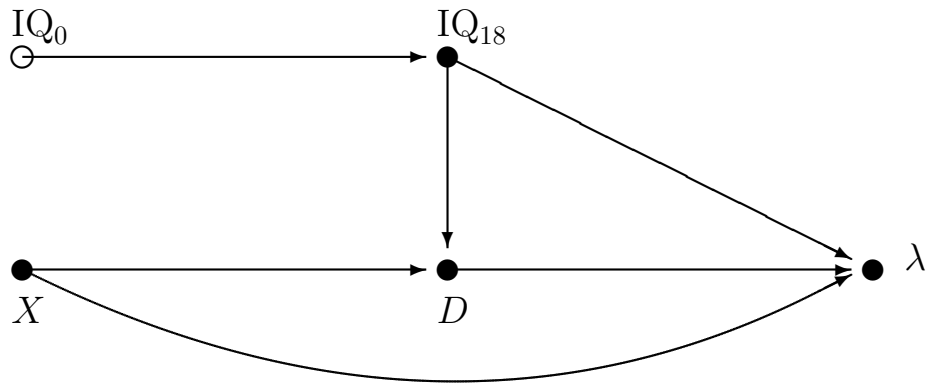


Figure 4: Directed acyclic graph of selection on IQ

We re-estimate the propensity scores, now including the IQ-test. The results confirm a large impact of the IQ on the education choice.[8] Based on these propensity scores we calculate, for each pairwise education comparison, the standardized bias, using (8). Table 6 shows the percentage bias measure before and after adjusting the data in our sample. Including IQ in the propensity score removed most of the imbalance in the values of the IQ-test between two adjacent education levels.

[7]The results when U follows the distribution of any of the other included binary variables are available upon request.
[8]See Table B.3 in Appendix B for the full results.

Table 6: Standardized bias before and after matching, pairwise comparisons (propensity score with IQ)

	Primary to lower vocational		Lower vocational to lower secondary		Lower secondary to higher	
	Before	After	Before	After	Before	After
Father's occupation						
Professional	4.9	1.1	20.2	−0.6	48.8	−0.7
self-employed	−0.9	0.7	5.4	0.3	−10.1	−0.8
Skilled	−9.8	0.5	−22.9	0.1	−38.9	2.9
Unskilled	−20.4	2.1	−17.0	0.1	−24.2	−1.0
Missing	−5.9	0.9	−4.8	−0.6	−6.1	1.1
Family size	7.6	−1.0	7.5	−0.4	−2.4	1.3
Born in Utrecht	−11.9	0.4	2.4	−0.2	3.8	−1.2
Religion						
Catholic	−16.1	−1.5	−4.9	−0.0	2.3	0.5
Dutch Reformed	12.6	1.8	0.7	0.1	−2.5	−2.6
Calvinist	17.4	−2.7	3.5	−0.4	2.2	−0.3
Other religion	−0.3	2.7	3.1	−0.3	2.4	1.2
Health						
Bad general health	−9.2	1.7	6.9	0.1	−1.0	−0.1
Bad hearing	−8.6	0.4	−5.1	−0.3	−2.2	−0.8
Bad sight	9.9	0.9	29.6	−0.8	21.6	−1.7
Bad psychological	−35.9	1.8	−2.2	−0.3	−3.9	1.1
Comprehensive IQ						
1 (highest)	−35.0	−4.8	40.8	−1.2	77.1	−0.6
2	69.3	−1.8	43.0	0.8	−20.9	−0.8
3	41.0	1.5	−21.7	0.5	−20.9	0.2
4	−12.9	1.1	−44.4	−0.2	−34.5	1.1
5	−70.8	1.0	−37.9	−0.7	−16.6	4.1
6 (lowest)	−45.5	0.3	−10.6	0.7	−3.5	−0.7
missing	−9.2	0.4	4.5	0.1	3.8	−0.3

Based on these new propensity scores we re-estimate the impact of education on the mortality rate, using a Cox model, a Gompertz model and Gompertz model with Gamma distributed unobserved heterogeneity. Table 7 presents the estimated effect on the mortality hazard of moving up one educational level for the Cox and the Gamma-Gompertz model. Accounting for selection on IQ removes most of the significant impact of education on mortality. Now, only men with primary education would gain from moving up the education ladder, with an 12% reduction in mortality.

Table 7: IPW-total effect including IQ-score in propensity score

	Cox			Gamma-Gompertz		
	ATE	ATT	ATU	ATE	ATT	ATU
Primary to	−0.026	0.024	-0.108^{+}	−0.026	0.025	-0.115^{+}
lower vocational	(0.056)	(0.070)	(0.050)	(0.056)	(0.070)	(0.055)
Lower vocational to	0.029	0.016	0.029	0.029	0.016	0.032
lower secondary	(0.035)	(0.037)	(0.039)	(0.035)	(0.037)	(0.042)
Lower secondary to	−0.091	−0.090	−0.088	−0.109	−0.104	−0.104
higher	(0.061)	(0.050)	(0.076)	(0.073)	(0.060)	(0.088)

$^{+}p < 0.05$ and $^{**}p < 0.01$

5.1 Sensitivity analyses

Throughout we have assumed that the propensity scores are estimated consistently. Misspecification of the propensity score will generally produce bias. An approach to improve the robustness of the proposed methodology can be obtained using a doubly robust estimator which also includes a regression adjustment. Rotnitzky and Robins (1995) point out that if either the regression adjustment or the propensity score is correctly specified the resulting estimator will be consistent. Thus we also estimate doubly robust estimators of the models, including the observed charateristics and the IQ-test both in the propensity score and in the hazard regression, see Table 8. Including regression covariates hardly changes the IPW estimates. Not surprisingly, including the covariates does change the "unadjusted" results.

Table 8: Robust estimation of education effect, including IQ

	Cox				Gamma-Gompertz			
	Unadjusted	IPW			Unadjusted	IPW		
		ATE	ATT	ATU		ATE	ATT	ATU
Primary to	-0.093^{+}	-0.026	0.021	-0.124^{+}	-0.101^{+}	-0.026	0.027	-0.135^{+}
lower voc.	(0.042)	(0.055)	(0.070)	(0.051)	(0.056)	(0.057)	(0.075)	(0.058)
Lower voc. to	0.029	0.029	0.019	0.036	0.033	0.035	0.022	0.043
lower sec.	(0.034)	(0.035)	(0.036)	(0.041)	(0.036)	(0.039)	(0.039)	(0.046)
Lower sec. to	-0.104^{+}	-0.085	-0.089	-0.081	-0.116^{+}	-0.107	-0.105	-0.105
higher	(0.047)	(0.063)	(0.051)	(0.078)	(0.053)	(0.073)	(0.063)	(0.090)

$^{+}p < 0.05$ and $^{**}p < 0.01$

Again we test the assumption of no selection on unmeasured confounding using the sensitivity analyses described in Section 4.1. The only difference is that we now also have the IQ-measures included in the base propensity score. The configurations of p_{ij} are chosen again to replicate the probability that a binary variable is equal to one conditional on $D = 0, 1$ and $\delta = 0, 1$ for each included binary variable. For each configuration we simulate U 100 times, calculate the outcome and selection effects and the implied IPW impact of education on the mortality rate. For all these calculation the value of the IQ-test is now also included. The outcome and selection effects (given in Table B.5 in Appendix B) are very similar to the outcome and selection effect in the model ignoring intelligence.

The implied IPW estimations using an IPW Cox model including the additional variable U (and intelligence) in Table 9 show that the impact of education on mortality still may change with two percentage point with respect to the baseline (not including U), but none of these changes are signif-

Table 9: Sensitivity of impact of education

	Primary to lower vocational			Lower vocational to lower secondary			Lower secondary to higher		
	ATE	ATT	ATU	ATE	ATT	ATU	ATE	ATT	ATU
IQ-test									
1 (highest)	0.040	0.098	−0.085	0.064	0.070	0.049	−0.019	−0.035	−0.011
	(0.322)	(0.401)	(0.052)	(0.038)	(0.044)	(0.042)	(0.091)	(0.070)	(0.118)
2	0.108	0.198	−0.079	0.050	0.043	0.046	−0.080	−0.098	−0.069
	(0.183)	(0.250)	(0.057)	(0.038)	(0.042)	(0.044)	(0.065)	(0.051)	(0.082)
4	−0.026	0.023	-0.104^{+}	0.062	0.029	0.079	−0.087	−0.068	−0.096
	(0.057)	(0.072)	(0.051)	(0.042)	(0.040)	(0.055)	(0.182)	(0.053)	(0.231)
5	0.031	0.079	−0.050	0.064	0.036	0.076	−0.095	−0.082	−0.098
	(0.079)	(0.104)	(0.080)	(0.048)	(0.039)	(0.070)	(0.076)	(0.051)	(0.100)
6 (lowest)	−0.004	0.044	−0.085	0.035	0.018	0.039	−0.091	−0.089	−0.088
	(0.075)	(0.080)	(0.126)	(0.039)	(0.037)	(0.049)	(0.061)	(0.050)	(0.076)

Based on adding U to propensity score with probabilities of U from observed for comprehensive IQ-test. Original estimates are ATE: -0.026 (Primary to lower voc.); 0.029 (Lower voc. to lower sec.); -0.091 (Lower sec. to higher); ATT: 0.024 (Primary to lower voc.); 0.016 (Lower voc. to lower sec.); -0.090 (Lower sec. to higher); ATU: -0.108 (Primary to lower voc.); 0.029 (Lower voc. to lower sec.); -0.088 (Lower sec. to higher).

icant. The average treatment on the untreated of education on the mortality rate for men with only primary education is, contrary to the baseline, only significant when U follows a distribution close to the middle IQ-values.

6 Mediation analysis for the mortality hazard rate

An alternative to assuming that intelligence at age 18 is a proxy for early childhood intelligence that should be accounted for in the education choice is to assume that education raises intelligence and that part of the impact of education on mortality runs through increased intelligence. In this section we discuss a model in which intelligence at age 18 mediates the impact of education on mortality. Mediation analysis aims to unravel the underlying causal mechanism into direct and indirect effects of an intermediate variable, the *mediator*. The counterfactual notation used in Section 3 for average treatment effects can be extended to define causal mediation, (see Huber 2014). We are particularly interested in the mediating effect of intelligence on mortality. It has been proven that high intelligence is positively associated with high education, (Ceci 1991; Hansen et al. 2004; Carlsson et al. 2015). Note that this does not rule out that early childhood intelligence influences the education choice. We use M_i to denote the observed intelligence level, which is measured around age 18 when the men had their military examination and after they had completed secondary schooling. The mediation model we assume is illustrated by the DAG in Figure 5

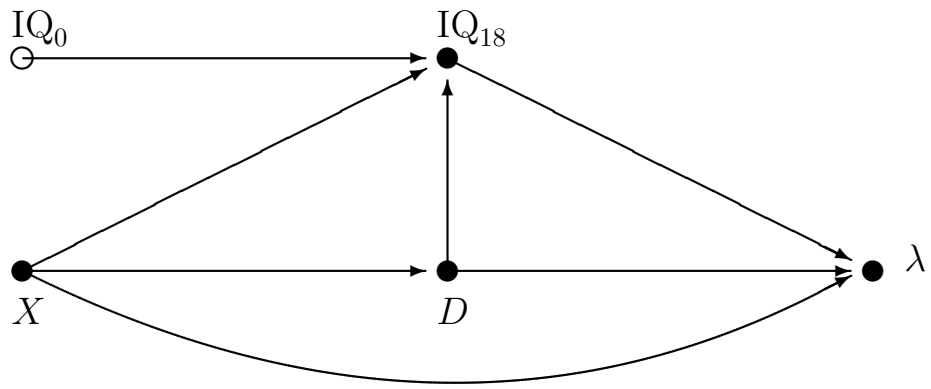


Figure 5: Directed acyclic graph of mediation through IQ_{18} conditional on X

Traditionally, causal mediation analysis has been formulated with the framework of linear structural models (Baron and Kenny 1986). Recent papers have placed causal mediation analysis within the counterfactual/potential outcomes framework (Imai et al. 2010; Imai et al. 2011; Huber 2014). Previously, the potential outcome was solely a function of the treatment, e.g. education choice, but in mediation analysis the potential outcomes also depend on the mediator. Because intelligence can

be affected by the education attained[9], there exist two potential values, $M_i(1)$ and $M_i(0)$, only one of which will be observed, i.e. $M_i = D_i \cdot M_i(1) + (1 - D_i) \cdot M_i(0)$. For example, if individual i actually attained education level 1, we would observe $M_i(1)$ but not $M_i(0)$. Next we use $\lambda_i\big(t|d, m\big)$ to denote the potential mortality hazard that would result from education equals d and intelligence equals m. For example, in the conscription data, $\lambda_i\big(t|1, 110\big)$ represents the mortality hazard that would have been observed if individual i had education level 1 and then had a measured IQ-score of 110. As before, we only observe one of the multiple hazards $\lambda_i = \lambda_i\big(t|D_i, M_i(D_i)\big)$.

Because we base our treatment effect on (mixed) proportional hazard models, it is natural to define the direct effect proportionally (just as the other treatment effects before). Abbring and van den Berg (2003) also define, in a different setting with a dynamic treatment, a proportional treatment effect for a duration outcome. In other non-linear settings, like count data, a proportional treatment effect has been defined (Lee and Kobayashi 2001). We define the (average) direct effect, depending on treatment status d :

$$\theta(d) = \frac{\mathrm{E}\Big[\lambda\big(t|1, M(d)\big)\Big]}{\mathrm{E}\Big[\lambda\big(t|0, M(d)\big)\Big]} \tag{11}$$

The framework we use enables us to disentangle the underlying causal pathway from education to mortality into a direct and an indirect effect of education through intelligence. The direct effect equals the causal effect of the treatment, moving up one education level, on the outcome, the mortality hazard that is not transmitted by the mediator, intelligence. First, we assume conditional independence (given X) of the treatment and the mediator:

Assumption 3. *Sequential ignorablility:* $\{\lambda(t|d', m), M(d)\} \perp D|X$ and $\lambda(t|d', m) \perp M|D = d, X$, $\forall d, d' = 0, 1$ and m in the support of M.

This implies that, conditional on observed covariates X, the treatment is jointly independent of the potential hazard and the mediator and that the observed mediator is ignorable given the value of the observed covariates and the treatment. Assumption 3 is a strong assumption and nonrefutable. We therefore carry out a set of sensitivity analyses to quantify the robustness of our empirical findings to violation of the sequential ignorability assumption, introducing an additional (unobserved) binary variable needed for the sequential ignorability to hold. This is an extension of the sensitivity analyses of the total IPW effect. As with the propensity score we also have a common support restriction:

[9]For example, Jones et al. (2011) discuss how performance in IQ tests could be influenced by coaching received by primary school pupils to prepare them for entrance tests for secondary school.

Assumption 4. *Common support mediator:* $0 < \Pr(D = 1|M, X) < 1$.

In addition we assume:

Assumption 5. *Proportional direct effect:* $\lambda\big(t|1, M(d)\big) = e^{\theta(d)}\lambda\big(t|0, M(d)\big)$.

This is equivalent to assuming that the effect of the treatment, D, is not moderated by the value of the mediator. Thus, we assume no interaction effect, $D \cdot M$, in the hazard. Note that this does not rule out an MPH model. It only assumes that the unobserved heterogeneity V is independent of the treatment D (as before) and the mediator M. Huber (2014) provides the same assumptions for identification of the direct and indirect effects in a 'standard' mean difference outcome model. This leads to the following identification theorem for the direct effect of a treatment on the hazard:

Theorem 1: Identification of direct effect $\theta(d)$.
Under Assumptions 1 to 5 the direct effect is identified through a weighted Cox or MPH regression with weights:

$$W(d) = \frac{\Pr(D = d|M, X)}{\Pr(D = d|X)}\left(\frac{D}{\Pr(D = 1|M, X)} + \frac{1 - D}{\Pr(D = 0|M, X)}\right) \tag{12}$$

with weight $W(d)$ for $\theta(d)$, for $d = 0, 1$.
(See Appendix A for the proof.)

The 'total effect' of education on the mortality rate, from an IPW estimation in which the mediator is excluded from the propensity score, can be decomposed into a direct effect of education and an indirect effect running through the mediator intelligence:

$$\frac{\lambda\big(t|D = 1, M(1)\big)}{\lambda\big(t|D = 0, M(0)\big)} = \frac{\lambda\big(t|D = 1, M(1)\big)}{\lambda\big(t|D = 0, M(1)\big)} \cdot \frac{\lambda\big(t|D = 0, M(1)\big)}{\lambda\big(t|D = 0, M(0)\big)} = \exp\Big(\theta(1) + \eta(0)\Big) \tag{13}$$

$$= \frac{\lambda\big(t|D = 1, M(1)\big)}{\lambda\big(t|D = 1, M(0)\big)} \cdot \frac{\lambda\big(t|D = 1, M(0)\big)}{\lambda\big(t|D = 0, M(0)\big)} = \exp\Big(\eta(1) + \theta(0)\Big) \tag{14}$$

The direct effect can be estimated solving (5), for a Cox model, or (7), for an MPH model, using $W(d)$ in (12) as weights. The indirect effects can be obtained from log-difference of the estimated total and the estimated direct effect, using (13) or (14). The direct effect represents the effect of education on the mortality hazard while holding intelligence constant at the level that would have been realized for chosen education level d. The indirect effects represents the effect on mortality if one changes intelligence from the value that would have been realized for education level 0 to the value that would have been observed for education level 1, while holding the education level at level d.

For estimation we use normalized versions of the sample implied by the weights in (12), such that the weights in either treatment or control groups add up to unity, as advocated earlier:

$$\tilde{W}_i(1) = \left[\frac{D_i}{\hat{p}(X_i)} \Big/ \sum_{j=1}^{N} \frac{D_j}{\hat{p}(X_j)}\right] + \left[\frac{(1-D_i)\hat{p}(X_i,M_i)}{\hat{p}(X_i)\big(1-\hat{p}(X_i,M_i)\big)} \Big/ \sum_{j=1}^{N} \frac{(1-D_j)\hat{p}(X_j,M_j)}{\hat{p}(X_j)\big(1-\hat{p}(X_j,M_j)\big)}\right] \quad (15)$$

$$\tilde{W}_i(0) = \left[\frac{D_i\big(1-\hat{p}(X_i,M_i)\big)}{\hat{p}(X_i,M_i)\big(1-\hat{p}(X_i)\big)} \Big/ \sum_{j=1}^{N} \frac{D_j\big(1-\hat{p}(X_j,M_j)\big)}{\hat{p}(X_j,M_j)\big(1-\hat{p}(X_j)\big)}\right] + \left[\frac{(1-D_i)}{1-\hat{p}(X_i)} \Big/ \sum_{j=1}^{N} \frac{(1-D_j)}{\big(1-\hat{p}(X_j)\big)}\right] \quad (16)$$

where $\hat{p}(X_i, M_i)$ denotes the estimate of the propensity score $\Pr(D = 1|X_i, M_i)$, which we estimate by probit specifications.

To estimate the average direct treatment effect on the treated (ATT), we need to weight the contribution of $W(1)$ by the propensity score $\hat{p}(X_i)$.[10] Similarly, if we want to estimate the average direct effect on the untreated (ATU), we reweight the contribution of $W(0)$ by one minus the propensity score. Note that the ATT and ATU weights for the direct effects are exactly the same as the ATT and ATU weights for the total IPW effect when including IQ in the propensity score.

A nice feature of Theorem 1 is that it is straightforward to implement, and only involves the the estimation of two propensity scores and plugging them in standard (mixed) proportional hazard estimation. No parametric restriction is imposed on the model of the mediator. Tchetgen Tchetgen (2013) also defines mediation analysis of the direct effect in (Cox) proportional hazard models. His method implies estimating a regression model for the mediator conditional on the treatment and pre-treatment covariates, $f(M|D, X)$, and it is much more difficult to formulate a suitable model for the mediator than for the propensity score.

6.1 Empirical results for mediation analysis

In Table 10 we present the direct effect and indirect effects of education on the mortality rate through IQ. The direct effect of education is only significant for the highest education group, about $\frac{2}{3}$ of the mortality reduction for men moving from lower secondary to higher education is attributable to a direct education effect. For the lowest education group the increase in intelligence induced by the additional education, the indirect effect of education through intelligence, is more important than the direct effect of education. For these low educated men the reduction in mortality when improving education is for 90% explained by this mediation effect of intelligence. When focussing on the treatment on the

[10] Of course, only $\theta(1)$ is relevant when interested in treatment on the treated and only $\theta(0)$ when interested in the treatment on the untreated.

untreated, the effect of improving education on the mortality hazard for those with the lower education level, the direct effect of education is only significant for the lowest education group and explains about 50% of the reduction. Note that theses direct ATT and ATU of education are exactly equal to the ATT and ATU in a selection model with the IQ-measures in the propensity score. Accounting for unobserved heterogeneity in the mortality hazard only affects the estimation for the highest education group, depicted in the second panel of Table 10.

Table 10: Direct and indirect impacts of education levels on the mortality rate

	Average treatment effect (ATE)				ATT	ATU
	direct effect		*indirect effect*		*direct effect*	
	$\theta(1)$	$\theta(0)$	$\eta(0)$	$\eta(1)$	$\theta(1)$	$\theta(0)$
Cox						
Primary to	-0.022	-0.070	-0.164^{+}	-0.116	0.024	-0.108^{+}
lower vocational	(0.068)	(0.048)	(0.078)	(0.062)	(0.070)	(0.051)
Lower vocational to	0.038	0.036	-0.090	-0.088	0.016	0.029
lower secondary	(0.035)	(0.038)	(0.047)	(0.049)	(0.037)	(0.039)
Lower secondary to	-0.124^{+}	-0.079	-0.066	-0.110	-0.090	-0.086
higher	(0.050)	(0.069)	(0.068)	(0.083)	(0.050)	(0.076)
Gamma-Gompertz						
Primary to	-0.021	-0.069	-0.164^{+}	-0.116	0.025	-0.108^{+}
lower vocational	(0.068)	(0.048)	(0.079)	(0.062)	(0.070)	(0.051)
Lower vocational to	0.038	0.037	-0.090	-0.089	0.016	0.032
lower secondary	(0.035)	(0.040)	(0.047)	(0.051)	(0.037)	(0.042)
Lower secondary to	-0.149^{+}	-0.094	-0.067	-0.123	-0.104	-0.104
higher	(0.061)	(0.081)	(0.080)	(0.097)	(0.060)	(0.088)

$^{+}p < 0.05$ and $^{**}p < 0.01$

Again to account for possible misspecification of the propensity scores we also estimate doubly robust estimators of the models, including the covariates both in the propensity score and in the hazard regression. Including regression covariates hardly changes the IPW estimates. The table with detailed results can be found in Appendix B.

6.2 Sensitivity analyses

For identification of the mediation effects we impose Assumption 3 of sequential ignorability. We extend the sensitivity analyses in Section 4.1 to assume that conditional on the binary (unobserved) factor the following two conditions hold (i) $\{\lambda(t|d',m), M(d)\}\perp D|X,U$ and (ii) $\lambda(t|d',m)\perp M|D = d,X,U$ for $\forall d,d' = 0,1$ and m in the support of M. These conditions lead to one of two following probabilities that define the distribution of U:

$$p^{\delta}_{mj} = \Pr(U=1|M=m,\delta=j,X) = \Pr(U=1|M=m,\delta=j) \quad (17)$$

$$p^{M}_{mi} = \Pr(U=1|M=m,D=i,X) = \Pr(U=1|M=m,D=i) \quad (18)$$

where $m = 1,\ldots,6$ the six possible values of the IQ-tests. Similar to the previous sensitivity analyses we define the outcome-effect as ω, with ω is the coefficient of U in a Cox model for the control group $(D=0)$ using U,X and M as covariates, the selection effect ξ, with ξ is the coefficient of U in a logit model on the treatment assignment (D) using U and X as covariates. A new measure, the mediator-effect, is ψ, with ψ is the coefficient of U in an ordered logit model on the IQ-test values for the control group using U and X as covariates.

The configurations of p^{δ}_{mj} and p^{M}_{mi} are chosen such that they mimic the probability that a binary variable is equal to one conditional on M and δ or M and D for each included binary variable. For each configuration we simulate U 100 times, calculate the outcome and selection effects and the implied IPW impact of education on the mortality rate. For all these calculation the value of the IQ-test is now also included. These outcome, selection and mediator effects are rather small and only a few are significantly different from zero. Next we re-estimate the direct IPW Cox including U in the propensity score. None of the assumed distributions of the unobserved confounder U leads to a substantial change in the estimated direct effects of education on mortality.[11]

7 Comparison intelligence as a selection and as a mediator

Table 11 summarises the results of all our analyses. It is clear that the unadjusted model largely overestimates the impact of education on mortality. When accounting for selective education choice and including intelligence at age 18 as one of the factors renders only the for the men with primary education improving education beneficial. When we view intelligence as a mediation factor in the causal chain from education to mortality and obtained the direct and indirect effects of education, we only find a significant direct effect if men with lower vocational education improved their education level to higher education, conditional on having the intelligence level of the higher educated. For

[11] The full tables of results can be found in Table B.8 to B.11 in Appendix B.

the men with only primary education we find the the indirect effect of education running through intelligence, improving the intelligence to the level of those with lower vocational education, improved the mortality.

Table 11: Comparison of impact of education on mortality for alternative models

	Primary to lower vocational	Lower vocational to lower secondary	Lower secondary to higher
Unadjusted	-0.236^{**}	-0.071^{+}	-0.220^{**}
	(0.035)	(0.030)	(0.041)
		IPW including IQ	
ATE	−0.026	0.029	−0.091
	(0.056)	(0.035)	(0.061)
ATT	0.024	0.016	−0.090
	(0.070)	(0.037)	(0.050)
ATU	-0.108^{+}	0.029	−0.086
	(0.051)	(0.039)	(0.076)
		IPW mediation	
ATE total	-0.185^{**}	−0.052	-0.190^{**}
	(0.039)	(0.031)	(0.046)
Direct effect			
ATE $\theta(1)$	−0.022	0.038	-0.124^{+}
	(0.068)	(0.035)	(0.050)
ATE $\theta(0)$	−0.070	0.036	−0.079
	(0.048)	(0.038)	(0.069)
ATT $\theta(1)$	0.024	0.016	−0.090
	(0.070)	(0.037)	(0.050)
ATU $\theta(0)$	-0.108^{+}	0.029	−0.086
	(0.051)	(0.039)	(0.076)
Indirect effect			
ATE $\eta(0)$	-0.164^{+}	−0.090	−0.066
	(0.078)	(0.047)	(0.068)
ATE $\eta(1)$	−0.116	−0.088	−0.110
	(0.062)	(0.049)	(0.083)

$^{+}p < 0.05$ and $^{**}p < 0.01$

From the Gompertz-hazards we can estimate the median survival age of the recruits and (slightly more complicated) the remaining life expectancy. The median survival age is the age at which half of the people have died (conditional on survival up to age 18, the age at the militarily examinations, when they were all alive). Assuming that the estimated Gompertz hazard holds for the remaining life, the life expectancy at age $t_0 = 18$ can be very well approximated by (see Lenart (2014)):

$$\mathrm{LE}(t_0) = -\exp\Big(e^{\alpha_0+\alpha_1 t_0}\Big)(\alpha_0 - \ln(\alpha_1) + \alpha_1 t_0 + 0.5772)/\alpha_1 \tag{19}$$

with 0.5772 is the Euler constant. For the unadjusted Gompertz model the estimated remaining life expectancy are 59.8 (Primary); 62.5 (lower vocational); 63.3 (lower secondary) (63.8 based on last two education groups) and 66.4 (higher). Leading to educational gains of 2.6, 0.8 and 2.6 in life expectancy. The median survival ages are 80.1 (Primary); 82.8 (lower vocational); 83.6 (lower secondary) (84.6) and 86.8 (higher). Thus leading to the same educational gains.

In Table 12 we only report the gains in life expectancy. Based on the IPW estimates with IQ included as a selection variable, the first panel of the table, we can conclude that if an individual had improved his education he would gain little in life-expectancy. From primary to lower vocational education he would have gained 0.3 additional years (and his median age also would have improved by 0.3 years). If an individual had improved from lower vocation to lower secondary the gain in life expectancy is negative,-0.3 years. The gain in life expectancy if an individual had improved his education from lower secondary to higher education is 1.1 years. If the men with primary education had increased their education to lower vocational they would have gained 1.3 additional years of living (ATU). For the other two groups the ATTs and ATUs are close to the ATEs.

The second panel of Table 12 reports the gains in life expectancy based on the mediation analysis and decompose these into a direct impact of education and an indirect impact through intelligence. Based on the IPW estimates we can conclude that if an individual had improved his education from primary to lower vocational education he would have gained 2.1 additional years (and his median age also would have improved by 2.1 years). However, only 0.6 years of this gain are attributable to the direct effect of education and 1.3 years to the indirect effect of intelligence (0.2 and 1.9 for those who have vocation education). If an individual had improved from lower vocation to lower secondary the gain in life expectancy is only 0.8 years (and the direct effect of education of those who attained lower secondary is even negative). The gain in life expectancy if an individual had improved his education from lower secondary to higher education is 2.2 years. For those who attained higher education this gain in life expectancy is mainly attributable to the direct effect of education (1.5 years), while for those with lower secondary education the direct effect is smaller (0.9 years) than the indirect effect through intelligence.

Table 12: Gain in life expectancy

	Primary to lower vocational	Lower vocational to lower secondary	Lower secondary to higher
Unadjusted	2.7	0.8	2.6
	IPW including IQ		
ATE	0.3	−0.3	1.1
ATT	−0.3	−0.2	1.1
ATU	1.3	−0.3	1.0
	IPW mediation		
ATE total	2.1	0.6	2.2
Direct effect			
ATE $\theta(1)$	0.2	−0.4	1.5
ATE $\theta(0)$	0.8	−0.4	0.9
ATT $\theta(1)$	−0.3	−0.2	1.1
ATU $\theta(0)$	1.3	−0.3	1.0
Indirect effect			
ATE $\eta(0)$	1.9	1.0	0.7
ATE $\eta(1)$	1.3	1.0	1.3

8 Discussion

A large literature documents that higher levels of education are positively associated with a longer life. Possible mechanisms include occupational risks, health behavior, the ability to process information and intelligence (Cutler and Lleras-Muney 2008). It is commonly acknowledged that education and intelligence are correlated. Intelligence may cause differences in educational outcomes or education may cause intelligence differences. Most of the economics literature on the causal effect of education on health focuses on accounting for endogenous selection into education due to confounding third factors, such as intelligence, either by exploiting natural experiments in education due to changes in compulsory schooling laws (Mazumder 2012) or by defining a structural model (Conti et al. 2010; Bijwaard et al. 2015). The estimates based on natural experiments find little to no effect of education on health, while the studies based on structural models find that around half of the difference in health by education is due to selection. An alternative perspective is that intelligence is part of the causal pathway to the effect on mortality. It has been proven that high scores on intelligence tests are positively associated with schooling level, (Ceci 1991; Hansen et al. 2004; Carlsson et al. 2015).

In this paper we show that different assumptions about the place of intelligence, measured at late adolescence, in the causal path from education to mortality hardly affects the estimated impact of education on mortality. We estimate and compare two models. In the first model we assume that intelligence at age 18 is a proxy for early childhood intelligence and is an important factor determining the education choice. In the second model we assume that this intelligence is affected by education attained and has a mediating effect on the mortality difference across education groups. For both models we developed an inverse probability weighting (IPW) method for hazard models to estimate the impact of education on the mortality rate. We use conscription data of Dutch men born in 1944-1947 who were examined for military service between 1961-1965, and linked to national death records, in which we identified four education groups. Using the IPW methods we estimate, for each adjacent education group, the impact of improving education on the mortality risk. In the first model we obtain the total impact of education on mortality, while in the second model we decompose the impact into a direct educational effect and an indirect effect running through intelligence.

The results show that accounting for intelligence, either as a selection factor or as a mediator factor leads to little educational gain in mortality. In the selection model, the only significant result we find is that men with only primary education would have reduced their mortality rate by 11% if they had improved their education to lower vocational (Average treatment on the untreated). This amounts to 1.3 additional years of life. When accounting for intelligence as a mediator the direct effect of education is only significant for highest education group (about 12% reduction in the mortality rate),

leading to 1.5 additional years of life. For men with primary school we find a significant indirect effect of intelligence, through education, on mortality risk (about 15% reduction in the mortality rate), which is equivalent to 1.9 years longer life expectancy.

A limitation of our data, based on military entrance examination, is that we only observe men and no information on women is available. Bijwaard et al. (2015) found that educational gains for women appear to be higher than for men, in spite of the higher survival difference of women with lower versus higher education. These findings are based on much smaller numbers than the current study however and therefore need to be interpreted with caution. Another issue is that in the 1960s a major change occurred in the education system in the Netherlands and some of the specific education strata in this study no longer exist. In addition, the percentage of people with more than six years of post-primary school education is currently much higher compared to the past. These changes are not likely to affect our general conclusion that increased education only has a small effect on survival, but further long term studies will be needed to quantify these effects for contemporary school types. The issue of reverse causality that early childhood health affects educational attainment might distort our analyses (Case et al. 2005; Currie 2009). We have no information about childhood health status, which prevents us from investigating the possibility of reverse causality from health to education in our sample. An advantage of the IPW-methods we apply in this paper is that they are easy to implement is standard statistical software packages such as STATA.

References

Aalen, O. O., O. Borgan, and H. K. Gjessing (2009). *Survival and Event History Analysis*. New York: Springer–Verlag.

Abbring, J. H. and G. J. van den Berg (2003). The non–parametric identification of treatment effects in duration models. *Econometrica 71*, 1491–1517.

Albouy, V. and L. Lequien (2009). Does compulsory education lower mortality? *Journal of Health Economics 28*(1), 155–168.

Andersen, P. K. and O. Borgan (1985). Counting process models for life history data: A review. *Scandinavian Journal of Statistics 12*, 97–158.

Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding (1993). *Statistical Models Based on Counting Processes*. New York: Springer–Verlag.

Auld, M. C. and N. Sidhu (2005). Schooling, cognitive ability and health. *Health Economics 14*(10), 1019–1034.

Austin, P. C. (2014). A tutorial on the use of propensity score methods with survival or time-to-event outcomes: Reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine 33*, 1242–1258.

Baron, R. M. and D. A. Kenny (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personlaity and Social Psychology 51*, 1173–1182.

Basu, A., A. M. Jones, and P. Rosa Dias (2014). Long-term effects of school quality on health and lifestyle: Evidence from comprehensive schooling reforms in England. Working paper no. 20811, NBER.

Batty, G. and G. David (2004). Early life intelligence and adult health. *British Medical Journal 329*(7466), 585–586.

Batty, G., G. David, I. J. Deary, and L. S. Gottfredson (2007). Premorbid (early life) IQ and later mortality risk: Systematic review. *Annals of Epidemiology 17*(4), 278–288.

Bijwaard, G. E. and H. van Kippersluis (2015). Efficiency of health investment: Education or intelligence? Discussion Paper 15–004, Tinbergen Institute.

Bijwaard, G. E., H. van Kippersluis, and J. Veenman (2015). Education and health: the role of cognitive ability. *Journal of Health Economics 42*, 29–43.

Bruggink, J. W. (2009). Ontwikkelingen in de (gezonde) levensverwachting naar opleidingsniveau. *Bevolkingstrends 57*(4), 71–74.

Caliendo, M. and S. Kopeinig (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys 22*, 31–72.

Calvin, C. M., I. J. Deary, C. Fenton, B. A. Roberts, G. Der, N. Leckenby, and G. D. Batty (2011). Intelligence in youth and all-cause-mortality: Systematic review with meta-analysis. *International Journal of Epidemiology 40*(3), 626–640.

Carlsson, M., G. B. Dahl, B. Öckert, and D.-O. Rooth (2015). The effect of schooling on cognitive skills. *The Review of Economics and Statistics 97*, forthcoming.

Case, A., A. Fertig, and C. Paxson (2005). The lasting impact of childhood health and circumstance. *Journal of Health Economics 24*, 365–389.

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? a reassessment of the evidence. *Developmental Psychology 27*, 703–722.

Clark, D. and H. Royer (2013). The effect of education on adult mortality and health: Evidence from Britain. *American Economic Review 103*(6), 2087–2120.

Cole, S. R. and M. A. Hernán (2004). Adjusted survival curves with inverse probability weights. *Computer Methods and Programs in Biomedicine 75*, 45–49.

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.

Conti, G., J. J. Heckman, and S. Urzua (2010). The education-health gradient. *American Economic Review 100*, 234–238.

Currie, J. (2009). Healthy, wealthy, and wise: Socioeconomics status, poor health in childhood, and human capital development. *Journal of Economic Literature 41*, 87–122.

Cutler, D. and A. Lleras-Muney (2008). Education and health: Evaluating theories and evidence. In J. S. House, R. F. Schoeni, G. A. Kaplan, and H. Pollack (Eds.), *Making Americans Healthier: Social and Economic Policy as Health Policy*. New-York: Russell Sage Foundation.

Deary, I. J. (2008). Why do intelligent people live longer? *Nature 456*, 175–176.

Deary, I. J. and W. Johnson (2010). Intelligence and education: causal perceptions drive analytic processes and therefore conclusions. *International Journal of Epidemiology 39*, 1362–1369.

Doob, L. J. (1953). *Stochastic Processes*. New York: Wiley.

Doornbos, G. and D. Kromhout (1990). Educational level and mortality in a 32-year follow-up study of 18-year-old men in the Netherlands. *International Journal of Epidemiology 19*, 374–379.

Ekamper, P., F. van Poppel, A. D. Stein, and L. H. Lumey (2014). Independent and additive association of prenatal famine exposure and intermediary life conditions with adult mortality between age 18–63 years. *Social Science & Medicine 119*, 232–239.

Feng, P., X.-H. Zhou, Q.-M. Zou, M.-Y. Fan, and Z.-S. Li (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine 31*, 681–697.

Gavrilov, L. A. and N. S. Gavrilova (1991). *The Biology of Life Span: A Quantitative Approach*. New York: Harwood Academic Publisher.

Hansen, K. T., J. J. Heckman, and K. J. Mullen (2004). The effect of schooling and ability on achievement test scores. *Journal of Econometrics 121*, 39–98.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*, 1161–1189.

Huber, M. (2014). Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics 29*, 920–943.

Ichino, A., F. Mealli, and T. Nannicini (2008). From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *Journal of Applied Econometrics 23*, 305–327.

Imai, K., L. Keele, and D. Tingley (2010). A general approach to causal mediation analysis. *Psychological Methods 15*, 309–334.

Imai, K., L. Keele, D. Tingley, and T. Yamamoto (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review 105*, 765–789.

Imai, K. and D. A. van Dyk (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association 99*, 854–866.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity. *The Review of Economics and Statistics 86*, 4–29.

Jones, A. M., N. Rice, and P. Rosa Dias (2011). Long-term effects of school quality on health and lifestyle: Evidence from comprehensive schooling reforms in England. *Journal of Human Capital 5*, 342–376.

Kalbfleisch, J. D. and R. L. Prentice (2002). *The Statistical Analysis of Failure Time Data, second edition*. John Wiley and Sons.

Klein, J. P. and M. L. Moeschberger (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer–Verlag.

Lechner, M. (2002). Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. *Journal of the Royal Statistical Society: Series A 165*, 59–82.

Lee, M.-J. and S. Kobayashi (2001). Proportional treatment effects for count response panel data: Effects of binary excercise on health care demand. *Health Economics 10*, 411–428.

Lenart, A. (2014). The moments of the Gompertz distribution and the maximum likelihood of its parameters. *Scandinavian Actuarial Journal of the Institute of Actuaries 3*, 255–277.

Mazumder, B. (2008). Does education improve health: A reexamination of the evidence from compulsory schooling laws. *Economic Perspectives 33*, 2–16.

Mazumder, B. (2012). The effects of education on health and mortality. *Nordic Economic Policy Review 1*, 261–301.

Meghir, C., M. Palme, and E. Simeonova (2013). Education, cognition and health: Evidence from a social experiment. Working paper no. 19002, NBER.

Meyer, P.-A. (1963). Decomposition of supermartingales: The uniqueness theorem. *Illinois Journal of Mathematics 7*, 1–17.

Nannicini, T. (2007). A simulation-based sensitivity analysis for matching estimators. *STATA Journal 7*, 334–350.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press.

Pearl, J. (2012). The mediation formula: A guide to the assessment of causal pathways in nonlinear models. In C. Berzuini, P. Dawid, and L. Bernardinelli (Eds.), *Causality: Statistical Perspectives and Applications*, pp. 151–179. Chichester: John Wiley.

Ravelli, G. P., Z. A. Stein, and M. W. Susser (1976). Obesity in young men after famine exposure in utero and early pregnancy. *New England Journal of Medicine 7*, 349–354.

Robins, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran and D. Berry (Eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pp. 95–133. New York: Springer.

Robins, J. M., M. A. Hernán, and B. Brumback (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology 11*, 550–560.

Robins, J. M. and A. Rotnitzky (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In N. Jewell, K. Dietz, and V. Farewell (Eds.), *AIDS Epidemiology– Methodological Issues*, pp. 297–331. Boston: Birkhäuser.

Rosenbaum, P. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Rotnitzky, A. and J. M. Robins (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika 82*, 805–820.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non–randomized studies. *Journal of Educational Psychology 66*, 688–701.

Schröder, H. and H. B. G. Ganzeboom (2014). Measuring and modelling level of education in European societies. *European Sociological Review 47*(9), 119–136.

Tchetgen Tchetgen, E. J. (2013). Inverse odds ratio-weighted estimation for causal mediation analysis. *Statistics in Medicine 32*, 4567–4580.

Therneau, T. and P. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer–Verlag.

Van den Berg, G. J. (2001). Duration models: Specification, identification, and multiple duration. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics, Volume V*, Chapter 55, pp. 3381–3460. Amsterdam: North–Holland.

Van Kippersluis, H., O. O'Donnell, and E. van Doorslaer (2011). Long run returns to education: Does schooling lead to an extended old age? *Journal of Human Resources 46*(4), 695–721.

Vrooman, J. C. and J. Dronkers (1986). Changing educational attainment processes: Some evidence from the Netherlands. *Sociology of Education 59*(2), 69–78.

Appendix A Counting processes and proofs

To prove the consistency and the properties of our estimation strategy we rely on counting process theory for duration models. In a Cox-model the waiting time to some event T has a conditional distribution given observed X and treatment D with hazard rate

$$\lambda(t|D,X,M) = \lambda_0(t)e^{\beta' X+\gamma D} \tag{A.1}$$

The cdf and pdf of the distribution of the duration T can be expressed as functions of the hazard rate. The counting process approach has increasingly become the standard framework for analyzing duration data and Andersen et al. (1993) have provided an excellent survey of this literature. Less technical surveys have been given by Klein and Moeschberger (1997), Therneau and Grambsch (2000), and Aalen et al. (2009). The main advantage of this framework is that it allows us to express the duration distribution as a regression model with an error term that is a martingale difference. Regression models with martingale difference errors are the basis for inference in time series models with dependent observations. Hence, it is not surprising that inference is much simplified by using a similar representation in duration models.

To start the discussion, we first introduce some notation. A counting process $\{N(t); t \geq 0\}$ is a stochastic process describing the number of events in the interval $[0,t]$ as time proceeds. The process contains only jumps of size $+1$. For single duration data, the event can only occur once, because the units are observed until the event occurs. Therefore we introduce the observation indicator $Y(t) = I(T \geq t)$ that equal to 1 if the unit is under observation at time t and zero after the event has occurred. The counting process is governed by its random intensity process $Y(t)\lambda(t)$, with $\lambda(t)$ is the hazard in (4). If we consider a small interval $(t-dt, t]$ of length dt, then $Y(t)\lambda(t)$ is the conditional probability that the increment $\mathrm{d}N(t) = N(t) - N(t-dt)$ jumps in that interval given all that has happened until just before t. By specifying the intensity as the product of this observation indicator and the hazard rate we effectively limit the number of occurrences of the event to one. It is essential that the observation indicator only depends on events up to time t.

Usually, some of the observations are right-censored $\tilde{T} = \min(T, C_r)$. By defining the observation indicator as the product of the indicator $I(t \leq T)$ and, if necessary, an indicator of the observation plan, we capture when a unit is at risk for the event. A related concept is *left-truncation*. Left truncation occurs when individuals are only observed conditional on survival up till some duration C_l, the age of military examination in our application. In the case of right censoring and left-truncation the at-risk indicator: $Y(t) = I(t \leq T)I(t \leq C_r)I(t \geq C_l)$. We assume that C_r, C_l and T are conditionally independent given X. The history up to t, $\overline{Y}(t)$ is assumed to be a left continuous function of t. The

history of the whole process also includes the (history of the) the covariates treatment and mediator. Thus, we have

$$\Pr\big(\mathrm{d}N(t) = 1|\overline{Y}(t), D, X, M\big) = Y(t)\lambda(t|X, D, M) \tag{A.2}$$

A fundamental result in the theory of counting processes, the *Doob-Meyer decomposition*[12], allows us to write

$$\mathrm{d}N(t) = Y(t)\lambda(t|\overline{X}(t)D, X, M)\mathrm{d}t + \mathrm{d}M(t) \tag{A.3}$$

with $M(t), t \geq 0$ a martingale with conditional mean and variance

$$\begin{aligned} \mathrm{E}\big(\mathrm{d}M(t)|\overline{Y}(t), D, X, M\big) &= 0 & \text{(A.4)} \\ \mathrm{Var}\big(\mathrm{d}M(t)|\overline{Y}(t), D, X, M\big) &= Y(t)\lambda(t|D, X, M)\mathrm{d}t & \text{(A.5)} \end{aligned}$$

The (conditional) mean and variance of the counting process are equal, so that the disturbances in (A.3) are heteroscedastic. The probability in (A.2) is 0, if the unit is no longer under observation. A counting process can be considered as a sequence of Bernoulli experiments, because if $\mathrm{d}t$ is small, (A.4) and (A.5) give the mean and variance of a Bernoulli random variable. The relation between the counting process and the sequence of Bernoulli experiments is given in (A.3), which can be considered as a regression model with an additive error that is a martingale difference. This equation resembles a time-series regression model. The Doob-Meier decomposition is the key to the derivation of the distribution of the estimators, because the asymptotic behavior of partial sums of martingales is well-known.

Note that the standard Cox model solves $\mathrm{E}\big[U(\theta)\big] = 0$ with

$$U(\theta) = \int \Big[X^* - \frac{\sum Y(t)X^* e^{\theta X^*}}{\sum Y(t) e^{\theta X^*}}\Big]\mathrm{d}N(t) \tag{A.6}$$

where $X^* = (X, D, M)'$ and $\theta = (\beta, \gamma, \alpha)'$.

Proof of equation (5)**: IPW Cox is unbiased**:
First we derive $\mathrm{E}\big[\sum Y(t)WDe^{\gamma D}\big]$, $\mathrm{E}\big[\sum Y(t)We^{\gamma D}\big]$, $\mathrm{E}\big[W\mathrm{d}N(t)\big]$, and $\mathrm{E}\big[WD\mathrm{d}N(t)\big]$. Redefine the propensity score $p(d) = \Pr(D_i = d|X_i)$, with $d = 0, 1$. Note that the integral of the sum is equal to the sum of the integrals.

[12] Doob (1953) published the Doob decomposition theorem which gives a unique decomposition for certain discrete time martingales. Meyer (1963) proved a continuous time version of the theorem, which became known as the Doob-Meyer decomposition. Both Andersen et al. (1993) and Aalen et al. (2009) provide a thorough discussion of this theorem.

$$\begin{aligned}
\mathrm{E}\big[\sum Y(t)WDe^{\gamma D}\big] &= \mathrm{E}\big[S(t|D,X)WDe^{\gamma D}\big] \\
&= \int \sum_d p(d)\mathrm{E}\Big[S(t|D=d,X=x)\frac{de^{\gamma d}}{p(d)}\Big]f_X(x)\mathrm{d}x \qquad \text{(A.7)} \\
&= e^{\gamma}S(t|D=1)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\big[\sum Y(t)We^{\gamma D}\big] &= \mathrm{E}\big[S(t|D,X)We^{\gamma D}\big] \\
&= \int \sum_d p(d)\mathrm{E}\Big[S(t|D=d,X=x)\frac{e^{\gamma d}}{p(d)}\Big]f_X(x)\mathrm{d}x \qquad \text{(A.8)} \\
&= e^{\gamma}S(t|D=1) + S(t|D=0)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\big[W\mathrm{d}N(t)\big] &= \mathrm{E}\big[\lambda(t|D,X)S(t|D,X)W\mathrm{d}t\big] \\
&= \int \sum_d p(d)\mathrm{E}\Big[f(t|D=d,X=x)\frac{1}{p(d)}\Big]\mathrm{d}t f_X(x)\mathrm{d}x \qquad \text{(A.9)} \\
&= \big[f(t|D=1) + f(t|D=0)\big]\mathrm{d}t
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\big[WD\mathrm{d}N(t)\big] &= \mathrm{E}\big[\lambda(t|D,X)S(t|D,X)WD\mathrm{d}t\big] \\
&= \int \sum_d p(d)\mathrm{E}\Big[f(t|D=d,X=x)\frac{d}{p(d)}\Big]\mathrm{d}t f_X(x)\mathrm{d}x \qquad \text{(A.10)} \\
&= f(t|D=1)\mathrm{d}t
\end{aligned}$$

This implies:

$$\begin{aligned}
\mathrm{E}\big[U(\gamma)\big] &= \int\Big\{\mathrm{E}\big[W\cdot D\mathrm{d}N(t)\big] - \mathrm{E}\big[W\mathrm{d}N(t)\big]\frac{\mathrm{E}\big[\sum Y(t)\cdot W\cdot De^{\gamma D}\big]}{\mathrm{E}\big[\sum Y(t)We^{\gamma D}\big]}\Big\} \\
&= \int \mathrm{d}t\Big\{f(t|D=1) - \big[f(t|D=1)+f(t|D=0)\big]\frac{e^{\gamma}S(t|D=1)}{e^{\gamma}S(t|D=1)+S(t|D=0)}\Big\} \\
&= \int \mathrm{d}t\Big\{f(t|D=1) - \big[f(t|D=1)+f(t|D=0)\big]\times \\
&\qquad \frac{e^{\gamma}\lambda(t|D=0)S(t|D=1)}{e^{\gamma}\lambda(t|D=0)S(t|D=1)+\lambda(t|D=0)S(t|D=0)}\Big\} \\
&= \int \mathrm{d}t\Big\{f(t|D=1) - \big[f(t|D=1)+f(t|D=0)\big]\frac{f(t|D=1)}{f(t|D=1)+f(t|D=0)}\Big\} = 0
\end{aligned}$$

In moving to the last line we assume $\lambda(t|D=1) = e^{\gamma}\lambda(t|D=0)$.

Proof of (7): IPW Gompertz is unbiased

In a parametric PH model the log-likelihood in counting process notation is (Andersen and Borgan 1985):

$$\ln L_i = \int \big[\ln \lambda_0(t_i;\alpha) + \gamma D_i\big]\, \mathrm{d}N(t_i) - \int_0^1 Y_i(s)\lambda_0(s;\alpha)e^{\gamma D_i}\, ds \tag{A.11}$$

where $\lambda_0(t;\alpha)$ is the baseline hazard with parameters α, e.g. for a Gompertz baseline hazard $\lambda_0(t;\alpha) = e^{\alpha_0+\alpha_1 t}$. Standard maximum likelihood estimation solves the roots of the derivatives of the log-likelihood:

$$U_\alpha(\theta) = \sum_{i=1}^N \Big[\int \frac{\partial\lambda_0(t_i;\alpha)/\partial\alpha}{\lambda_0(t_i;\alpha)}\, \mathrm{d}N(t_i) - \int_0^1 Y_i(s)\tfrac{\partial\lambda_0(s;\alpha)}{\partial\alpha}e^{\gamma D_i}\, \mathrm{d}s\Big] \tag{A.12}$$

$$U_\gamma(\theta) = \sum_{i=1}^N \Big[\int D_i\, \mathrm{d}N(t_i) - D_i\int_0^1 Y_i(s)\lambda_0(s;\alpha)e^{\gamma D_i}\, \mathrm{d}s\Big] \tag{A.13}$$

with $\theta = (\alpha,\gamma)'$ and $U_\alpha(\theta)$ and $U_\gamma(\theta)$ are the gradients of the log-likelihood w.r.t. α and γ. The IPW version includes the weights W in equation (A.12) and (A.13). Because our main parameter of interest is γ we only focus on $U_\gamma(\theta)$. To proof (7) we use similar reasoning as above. First, we derive $\mathrm{E}[WD\mathrm{d}N(t)]$ and $\mathrm{E}[\sum Y(t)\lambda_0(t;\alpha)e^{\gamma D}WD]$.

$$\begin{aligned}
\mathrm{E}\big[\sum Y(t)W\lambda_0(t;\alpha)e^{\gamma D}WD\big] &= \mathrm{E}\big[\lambda_0(t;\alpha)e^{\gamma D}S(t|D,X)WD\big] \\
&= \int\sum_d p(d)\mathrm{E}\Big[\lambda_0(t;\alpha)e^{\gamma D}S(t|D=d,X=x)\frac{de^{\gamma d}}{p(d)}\Big]f_X(x)\mathrm{d}x \\
&= e^{\gamma}\lambda_0(t;\alpha)S(t|D=1) = f(t|D=1)
\end{aligned} \tag{A.14}$$

From (A.10) we have $\mathrm{E}[WD\mathrm{d}N(t)] = f(t|D=1)\mathrm{d}t$. Thus, if we assume the right parametric model this implies that $U_\gamma(\theta)$ has zero mean.

Proof of (7): IPW (gamma)-Gompertz is unbiased

In a MPH model with a parametric baseline hazard and a unit-mean Gamma-distributed unobserved heterogeneity with variance σ^2 the (unconditional) hazard is:

$$\lambda(t|D) = \frac{\lambda_0(t;\alpha)e^{\gamma D}}{1+\sigma^2\int_0^t \lambda_0(s;\alpha)e^{\gamma D}\, ds}$$

and the likelihood (in counting process notation) is:

$$L_i = \Big[\frac{\lambda_0(t;\alpha)e^{\gamma D}}{1+\sigma^2\int Y_i(s)\lambda_0(s;\alpha)e^{\gamma D}\, ds}\Big]^{\mathrm{d}N_i(t)}\Big[1+\sigma^2\int Y_i(s)\lambda_0(s;\alpha)e^{\gamma D}\, ds\Big]^{-1/\sigma^2} \tag{A.15}$$

IPW solves the roots of the weighted derivatives of the log-likelihood. The weighted derivative w.r.t. γ is:

$$U_\gamma(\theta) = \sum_{i=1}^N\Big[\int \frac{W_iD_i}{1+\sigma^2\int Y_i(s)\lambda_0(s;\alpha)e^{\gamma D}\, ds}\mathrm{d}N_i(t) - \frac{W_iD_i\int Y_i(t)\lambda_0(t;\alpha)e^{\gamma D}\, dt}{1+\sigma^2\int Y_i(s)\lambda_0(s;\alpha)e^{\gamma D}\, ds}\Big] \tag{A.16}$$

To prove (7) we use similar reasoning as above. First, we derive

$$\begin{aligned}
\mathrm{E}\big[\frac{WD}{1+\sigma^2\int Y(s)\lambda_0(s;\alpha)e^{\gamma D}\,ds}\mathrm{d}N(t)\big] &= \mathrm{E}\Big[\frac{WD\lambda_0(t;\alpha)e^{\gamma D}Y(t)dt}{1+\sigma^2\int_0^t \lambda_0(s;\alpha)e^{\gamma D}\,ds}\Big] \qquad (A.17)\\
&= \int\sum_d p(d)\mathrm{E}\Big[\frac{\lambda_0(t;\alpha)e^{\gamma D}S(t|D=d,X=x)}{1+\sigma^2\int_0^t \lambda_0(s;\alpha)e^{\gamma D}\,ds}\frac{de^{\gamma d}}{p(d)}\Big]\mathrm{d}tf_X(x)\mathrm{d}x\\
&= \frac{\lambda_0(t;\alpha)e^{\gamma D}}{1+\sigma^2\int_0^t \lambda_0(s;\alpha)e^{\gamma D}\,ds}S(t|D=1)\mathrm{d}t = f(t|D=1)\mathrm{d}t
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\big[\sum\frac{WDY(t)\lambda_0(t;\alpha)e^{\gamma D}}{1+\sigma^2\int Y(s)\lambda_0(s;\alpha)e^{\gamma D}\,ds}\big] &= \qquad (A.18)\\
&= \int\sum_d p(d)\mathrm{E}\Big[\frac{\lambda_0(t;\alpha)e^{\gamma D}S(t|D=d,X=x)}{1+\sigma^2\int_0^t \lambda_0(s;\alpha)e^{\gamma D}\,ds}\frac{de^{\gamma d}}{p(d)}\Big]f_X(x)\mathrm{d}x\\
&= f(t|D=1)
\end{aligned}$$

Thus, if we assume the right parametric model for the baseline hazard and a Gamma distribution for the unobserved heterogeneity (A.17) has mean zero.[13]

Proof Theorem 1 and equation (12) **for Cox PH**[14]:

The direct effect $\theta(d)$ solves $\mathrm{E}\big[U(\theta(d))\big]=0$ with $U(\theta(d))$

$$U(\theta(d)) = \int W(d)\Big[D - \frac{\sum Y(t)W(d)De^{\gamma D}}{\sum Y(t)W(d)e^{\theta D}}\Big]\mathrm{d}N(t) \qquad (A.19)$$

Again we first derive $\mathrm{E}\big[\sum Y(t)W(d)De^{\theta D}\big]$, $\mathrm{E}\big[\sum Y(t)W(d)e^{\theta D}\big]$, $\mathrm{E}\big[W(d)\mathrm{d}N(t)\big]$, and $\mathrm{E}\big[W(d)D\mathrm{d}N(t)\big]$.

$$\begin{aligned}
\mathrm{E}\big[\sum Y(t)W(d)De^{\theta D}\big] &= \mathrm{E}\big[S(t|D,M,X)\cdot W(d)\cdot De^{\theta(d)D}\big]\\
&= \int e^{\theta(d)}\mathrm{E}\Big[S(t|D=1,M=m,X=x)\frac{\Pr(D=d|M,X)f_M(m|x)}{\Pr(D=d|X)}\Big]\mathrm{d}mf_X(x)\mathrm{d}x\\
&= \int e^{\theta(d)}\mathrm{E}\Big[S(t|D=1,M=m,X=x)f_M(m|D=d,X)\Big]\mathrm{d}mf_X(x)\mathrm{d}x\\
&= e^{\theta(d)}S\big(t|D=1,M(d)\big)
\end{aligned}$$

From line two to three we use Bayes' rule.

$$\begin{aligned}
\mathrm{E}\big[\sum Y(t)W(d)e^{\theta(d)D}\big] &= \mathrm{E}\big[S(t|D,M,X)\cdot W(d)e^{\theta(d)D}\big]\\
&= \int\sum_q \mathrm{E}\Big[S(t|D=q,M=m,X=x)f_M(m|D=d,X)\Big]\mathrm{d}mf_X(x)\mathrm{d}x\\
&= e^{\theta(d)}S\big(t|D=1,M(d)\big) + S\big(t|D=0,M(d)\big)
\end{aligned}$$

[13]The proof for any other MPH model with known functional form of the baseline hazard and given distribution of the unobserved heterogeneity is essentially the same.

[14]The proofs for the Gompertz PH and the Gamma-Gompertz MPH model are very similar and not shown here.

and

$$\begin{aligned}
\mathrm{E}\big[W(d)\mathrm{d}N(t)\big] &= \mathrm{E}\big[\lambda(t|D,M,X)S(t|D,M,X)\cdot W(d)\mathrm{d}t\big] \\
&= \int\sum_q \mathrm{E}\Big[f(t|D=q,M=m,X=x)f_M(m|D=d,X)\Big]\mathrm{d}t\mathrm{d}m f_X(x)\mathrm{d}x \\
&= \Big[f\big(t|D=1,M(d)\big)+f\big(t|D=0,M(d)\big)\Big]\mathrm{d}t
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\big[W(d)D\mathrm{d}N(t)\big] &= \mathrm{E}\big[\lambda(t|D,M,X)S(t|D,M,X)\cdot W(d)\cdot D\mathrm{d}t\big] \\
&= \int \mathrm{E}\Big[f(t|D=1,M=m,X=x)f_M(m|D=d,X)\Big]\mathrm{d}t f_X(x)\mathrm{d}x \\
&= f(t|D=1,M(d))\mathrm{d}t
\end{aligned}$$

This implies:

$$\begin{aligned}
\mathrm{E}\big[U(\theta(d))\big] &= \int\Big\{\mathrm{E}\big[W(d)\cdot D\mathrm{d}N(t)\big]-\mathrm{E}\big[W(d)\mathrm{d}N(t)\big]\frac{\mathrm{E}\big[\sum Y(t)\cdot W(d)\cdot De^{\theta(d)D}\big]}{\mathrm{E}\big[\sum Y(t)\cdot W(d)e^{\theta(d)D}\big]}\Big\} \\
&= \int \mathrm{d}t\Big\{f\big(t|D=1,M(d)\big)-\Big[f\big(t|D=1,M(d)\big)+f\big(t|D=0,M(d)\big)\Big]\times \\
&\qquad \frac{e^{\theta(d)}S\big(t|D=1,M(d)\big)}{e^{\theta(d)}S\big(t|D=1,M(d)\big)+S\big(t|D=0,M(d)\big)}\Big\} \\
&= \int \mathrm{d}t\Big\{f\big(t|D=1,M(d)\big)-\big[f\big(t|D=1,M(d)\big)+f\big(t|D=0,M(d)\big)\big]\times \\
&\qquad \frac{e^{\theta(d)}\lambda\big(t|D=0,M(d)\big)S\big(t|D=1,M(d)\big)}{e^{\theta(d)}\lambda\big(t|D=0,M(d)\big)S\big(t|D=1,M(d)\big)+\lambda\big(t|D=0,M(d)\big)S\big(t|D=0,M(d)\big)}\Big\} \\
&= \int \mathrm{d}t\Big\{f\big(t|D=1,M(d)\big)-\big[f\big(t|D=1,M(d)\big)+f\big(t|D=0,M(d)\big)\big]\times \\
&\qquad \frac{f\big(t|D=1,M(d)\big)}{f\big(t|D=1,M(d)\big)+f\big(t|D=0,M(d)\big)}\Big\}=0
\end{aligned}$$

Appendix B Additional tables and figures

Table B.1: Ordered Probit estimates of IQ-level, by levels of education

	Primary to lower vocational	lower vocational to lower secondary	Lower secondary to higher
Education:			
Lower vocational	1.100**	–	–
Lower secondary	–	0.692**	–
Higher	–	–	0.872**
Father's occupation:			
Professional	0.012	0.010	−0.047+
Self-employed	−0.096**	−0.175**	−0.185**
Clerical	–	–	–
Skilled	−0.202**	−0.184**	−0.119**
Unskilled	−0.291**	−0.291**	−0.257**
Missing	−0.302**	−0.249**	−0.168**
Family size	0.080**	0.013	−0.138**
Born in Utrecht	−0.137**	−0.117**	−0.093**
Religion:			
Catholic	0.039+	0.014	−0.016
Dutch Reformed	0.006	0.013	0.005
Calvinist	0.210**	0.179**	0.110**
Other religion	0.062	0.086	−0.233**
None	–	–	–
Health:			
Bad general health	−0.102**	−0.114**	−0.163**
Bad hearing	−0.504**	−0.539**	−0.538**
Bad sight	0.221**	0.197**	0.134**
Bad psychological	−0.383**	−0.294**	−0.259**
Famine cohorts:			
A1	0.102**	0.084+	0.143**
A2	0.204**	0.094+	0.124**
B1	0.167**	0.141**	0.158**
B2	0.178**	0.149**	0.172**
D1	0.161**	0.145**	0.179**
D2	0.190**	0.140**	0.113**

$^{+}p < 0.05$ and $^{**}p < 0.01$

Famine cohorts from Ravelli et al. (1976): born in 7 cities in the West of Netherlands and A1: born Jan 1944- May 1944; A2: born Jun 1944- Oct 1944; B1: born Nov 1944- Jan 1944; B2: born Feb 1945- May 1945; D1: born Jun 1945- Sep 1945; D2: born Oct 1945- Dec 1945.

Table B.2: Probit estimates of propensity scores ignoring IQ, pairwise comparisons

	Primary to lower vocational	Lower vocational to lower secondary	Lower secondary to higher
Father's occupation:			
Professional	-0.221^{**}	0.123^{**}	0.478^{**}
Self-employed	-0.386^{**}	-0.030	-0.276^{**}
Clerical	–	–	–
Skilled	-0.404^{**}	-0.413^{**}	-0.581^{**}
Unskilled	-0.581^{**}	-0.460^{**}	-0.616^{**}
Missing	-0.443^{**}	-0.334^{**}	-0.205^{**}
Family size	0.260^{**}	0.151^{**}	0.001
Born in Utrecht	-0.252^{**}	0.057	0.082^{+}
Religion:			
Catholic	-0.089^{**}	-0.037	0.043
Dutch Reformed	0.154^{**}	-0.002	-0.040
Calvinist	0.403^{**}	-0.068^{+}	0.048
Other religion	0.136	0.232^{+}	0.086
None	–	–	–
Health:			
Bad general health	-0.066^{+}	0.109^{**}	-0.034
Bad hearing	-0.192^{**}	-0.213^{**}	-0.060
Bad sight	0.139^{**}	0.374^{**}	0.267^{**}
Bad psychological	-0.448^{**}	-0.061^{**}	-0.077^{**}
Famine cohorts:			
A1	0.199^{**}	0.052	-0.028
A2	0.117^{+}	0.035	-0.062
B1	0.105^{**}	0.099^{**}	0.020
B2	0.071^{+}	0.058^{+}	-0.058^{+}
D1	-0.017	0.026	-0.088^{+}
D2	0.008	0.099^{**}	0.049

$^{+}p < 0.05$ and $^{**}p < 0.01$

Famine cohorts from Ravelli et al. (1976): born in 7 cities in the West of Netherlands and A1: born Jan 1944- May 1944; A2: born Jun 1944- Oct 1944; B1: born Nov 1944- Jan 1944; B2: born Feb 1945- May 1945; D1: born Jun 1945- Sep 1945; D2: born Oct 1945- Dec 1945.

Table B.3: Probit estimates of propensity scores, pairwise comparisons including IQ

	Primary to lower vocational	Lower vocational to lower secondary	Lower secondary to higher
Father's occupation:			
Professional	-0.196^{**}	0.110^{**}	0.485^{**}
Self-employed	-0.333^{**}	0.051	-0.175^{**}
Clerical	–	–	–
Skilled	-0.263^{**}	-0.319^{**}	-0.511^{**}
Unskilled	-0.371^{**}	-0.310^{**}	-0.481^{**}
Missing	-0.250^{**}	-0.231^{**}	-0.141^{**}
Family size	0.207^{**}	0.149^{**}	0.074^{+}
Born in Utrecht	-0.135^{**}	0.125^{**}	0.139^{**}
Religion:			
Catholic	-0.135^{**}	-0.047^{+}	0.066^{+}
Dutch Reformed	0.143^{**}	-0.011	-0.041
Calvinist	0.278^{**}	-0.009	0.008
Other religion	0.053	0.232^{+}	0.259^{+}
None	–	–	–
Health:			
Bad general health	-0.060^{+}	0.111^{**}	-0.021
Bad hearing	-0.041	-0.147^{**}	-0.008
Bad sight	-0.042	0.244^{**}	0.169^{**}
Bad psychological	-0.154^{**}	0.068^{**}	0.047
Famine cohorts:			
A1	0.119	0.041	-0.049
A2	-0.003	0.019	-0.089
B1	0.010^{**}	0.064^{+}	0.010
B2	-0.030^{+}	0.012	-0.107^{**}
D1	-0.114^{**}	-0.020	-0.146^{**}
D2	-0.127^{+}	0.047	0.024
Comprehensive IQ			
1 (highest)	1.317^{**}	0.826^{**}	1.502^{**}
2	0.602^{**}	0.527^{**}	0.763^{**}
3	–	–	–
4	-0.590^{**}	-0.450^{**}	-0.420^{**}
5	-1.254^{**}	-0.835^{**}	-0.429^{+}
6 (lowest)	-1.985^{**}	-1.017^{**}	0.112
missing	-0.759^{**}	0.302^{**}	0.970^{**}

$^{+}p < 0.05$ and $^{**}p < 0.01$

Table B.4: Estimated coefficients included variables in Gamma-Gompertz robust (unadjusted) estimation

	Primary to lower vocational	Lower vocational to lower secondary	Lower secondary to higher
Father's occupation:			
Professional	−0.026	0.001	0.030
Self-employed	0.160^{+}	0.124^{+}	0.136
Clerical	–	–	–
Skilled	−0.016	0.025	0.159**
Unskilled	0.030	0.043	0.165^{+}
Missing	0.201**	0.184**	0.123
Family size	0.180^{+}	0.107^{+}	0.100
Born in Utrecht	0.281**	0.194**	0.102
Religion:			
Catholic	−0.095^{+}	−0.096^{+}	−0.029
Dutch Reformed	−0.071	−0.041	−0.034
Calvinist	−0.077	−0.116	−0.217**
Other religion	−0.285	−0.222	−0.399
None	–	–	–
Health:			
Bad general health	0.094	0.173**	0.254**
Bad hearing	0.342**	0.195^{+}	0.114
Bad sight	−0.041	−0.047	−0.023
Bad psychological	0.170**	0.134**	0.109^{+}
Scale (constant)	−10.011**	−10.091**	−10.155**
Shape (age)	0.091**	0.088**	0.088**
Gamma-var	0.754**	0.235**	0.932**

$^{+}p < 0.05$ and $^{**}p < 0.01$

Table B.5: Sensitivity analysis propensity score with IQ: outcome and selection effects (additional covariates)

	Primary to lower vocational		Lower vocational to lower secondary		Lower secondary to higher	
	ω	ξ	ω	ξ	ω	ξ
neutral	0	0	0	0	0	0
Professional	-0.188	0.168	-0.005	0.601^{**}	-0.075	1.105^{**}
	(0.163)	(0.095)	(0.097)	(0.059)	(0.088)	(0.061)
Self-employed	0.180	-0.037	0.118	0.220^{+}	0.051	-0.426^{**}
	(0.160)	(0.113)	(0.114)	(0.078)	(0.120)	(0.111)
Skilled	0.023	-0.200^{**}	-0.085	-0.506^{**}	0.111	-1.105^{**}
	(0.087)	(0.058)	(0.059)	(0.043)	(0.075)	(0.080)
Unskilled	0.071	-0.520^{**}	-0.023	-0.525^{**}	0.082	-1.056^{**}
	(0.105)	(0.068)	(0.084)	(0.061)	(0.101)	(0.127)
Missing	0.050	-0.227^{+}	0.282^{+}	-0.200^{+}	0.031	-0.308^{+}
	(0.155)	(0.104)	(0.107)	(0.087)	(0.144)	(0.138)
Born in Utrecht	0.247	-0.411^{**}	0.236^{+}	0.094	0.091	0.132
	(0.131)	(0.092)	(0.102)	(0.074)	(0.114)	(0.094)
Catholic	0.044	-0.334^{**}	-0.105	-0.104^{+}	0.017	0.047
	(0.089)	(0.058)	(0.061)	(0.042)	(0.069)	(0.055)
Dutch Reformed	-0.052	0.284^{**}	0.037	0.020	0.337	-0.057
	(0.102)	(0.060)	(0.060)	(0.042)	(0.070)	(0.055)
Calvinist	-0.304	0.791^{**}	0.018	0.131	-0.207	0.069
	(0.274)	(0.133)	(0.109)	(0.073)	(0.122)	(0.093)
Other religion	-0.270	-0.005	-0.086	0.359	-0.184	0.229
	(0.653)	(0.377)	(0.444)	(0.248)	(0.408)	(0.284)
Bad general health	0.082	-0.236^{**}	0.131	0.188^{**}	0.234^{**}	-0.022
	(0.111)	(0.071)	(0.073)	(0.051)	(0.079)	(0.066)
Bad hearing	0.363	-0.455^{**}	0.278	-0.307^{+}	0.143	-0.191
	(0.182)	(0.146)	(0.158)	(0.126)	(0.224)	(0.196)
Bad sight	-0.008	0.235^{**}	-0.030	0.639^{**}	-0.032	0.436^{**}
	(0.109)	(0.066)	(0.066)	(0.044)	(0.065)	(0.051)
Bad psychological	0.215^{+}	-0.771^{**}	0.137^{+}	-0.048	0.184^{+}	-0.099
	(0.088)	(0.063)	(0.067)	(0.048)	(0.077)	(0.064)

Based on adding U to propensity score with probabilities of U from observed probabilities of covariates. No effect would give $\omega = 0$ and $\xi = 0$. $^{+}p < 0.05$ and $^{**}p < 0.01$

Table B.6: Sensitivity analysis IPW-total with IQ-measure in propensity score (additional parameters)

	Primary to lower vocational			Lower vocational to lower secondary			Lower secondary to higher		
	ATE	ATT	ATU	ATE	ATT	ATU	ATE	ATT	ATU
neutral	−0.026	0.025	-0.109^{+}	0.029	0.016	0.030	−0.088	−0.089	−0.084
	(0.056)	(0.070)	(0.050)	(0.035)	(0.037)	(0.039)	(0.062)	(0.051)	(0.076)
Professional	−0.024	0.027	-0.108^{+}	0.032	0.016	0.034	−0.087	−0.075	−0.089
	(0.056)	(0.071)	(0.051)	(0.036)	(0.038)	(0.040)	(0.074)	(0.059)	(0.095)
Self-employed	−0.026	0.025	-0.108^{+}	0.028	0.014	0.029	−0.089	−0.088	−0.086
	(0.056)	(0.070)	(0.050)	(0.035)	(0.037)	(0.039)	(0.063)	(0.051)	(0.078)
Skilled	−0.026	0.027	-0.112^{+}	0.031	0.008	0.041	−0.067	−0.073	−0.063
	(0.056)	(0.071)	(0.051)	(0.036)	(0.038)	(0.041)	(0.088)	(0.054)	(0.116)
Unskilled	−0.021	0.032	-0.111^{+}	0.031	0.015	0.034	−0.077	−0.084	−0.071
	(0.058)	(0.075)	(0.052)	(0.036)	(0.037)	(0.041)	(0.076)	(0.052)	(0.099)
Missing	−0.025	0.025	-0.104^{+}	0.030	0.020	0.029	−0.088	−0.089	−0.084
	(0.056)	(0.071)	(0.051)	(0.035)	(0.037)	(0.039)	(0.062)	(0.051)	(0.077)
Born in Utrecht	−0.017	0.035	-0.101^{+}	0.028	0.014	0.029	−0.092	−0.091	−0.083
	(0.056)	(0.072)	(0.051)	(0.035)	(0.037)	(0.039)	(0.062)	(0.051)	(0.076)
Catholic	−0.024	0.031	-0.116^{+}	0.028	0.014	0.030	−0.092	−0.090	−0.089
	(0.057)	(0.073)	(0.051)	(0.035)	(0.037)	(0.039)	(0.061)	(0.050)	(0.076)
Dutch Reformed	−0.024	0.028	-0.110^{+}	0.028	0.016	0.029	−0.091	−0.089	−0.088
	(0.057)	(0.072)	(0.051)	(0.035)	(0.037)	(0.039)	(0.062)	(0.050)	(0.076)
Calvinist	−0.017	0.037	-0.108^{+}	0.030	0.016	0.031	−0.090	−0.089	−0.087
	(0.059)	(0.076)	(0.051)	(0.035)	(0.037)	(0.039)	(0.062)	(0.050)	(0.076)
Other religion	−0.026	0.025	-0.108^{+}	0.029	0.016	0.030	−0.090	−0.090	−0.087
	(0.056)	(0.070)	(0.050)	(0.035)	(0.037)	(0.039)	(0.062)	(0.050)	(0.076)
Bad general health	−0.023	0.028	-0.104^{+}	0.023	0.012	0.022	−0.090	−0.089	−0.088
	(0.056)	(0.071)	(0.051)	(0.035)	(0.037)	(0.039)	(0.061)	(0.050)	(0.076)
Bad hearing	−0.020	0.032	-0.103^{+}	0.030	0.018	0.030	−0.090	−0.089	−0.087
	(0.056)	(0.071)	(0.051)	(0.035)	(0.037)	(0.039)	(0.062)	(0.050)	(0.076)
Bad sight	−0.025	0.025	-0.107^{+}	0.033	0.020	0.034	−0.097	−0.086	−0.098
	(0.056)	(0.072)	(0.051)	(0.036)	(0.039)	(0.042)	(0.065)	(0.052)	(0.082)
Bad psychological	0.012	0.070	−0.087	0.030	0.017	0.031	−0.090	−0.087	−0.088
	(0.063)	(0.083)	(0.056)	(0.035)	(0.037)	(0.039)	(0.062)	(0.051)	(0.076)

Based on adding U to propensity score with probabilities of U from observed probabilities for observed variables.
Original estimates are ATE: -0.026 (Primary to lower voc.); 0.029 (Lower voc. to lower sec.); -0.091 (Lower sec. to higher); ATT: 0.024 (Primary to lower voc.); 0.016 (Lower voc. to lower sec.); -0.090 (Lower sec. to higher); ATU: -0.108 (Primary to lower voc.); 0.029 (Lower voc. to lower sec.); -0.088 (Lower sec. to higher).

Table B.7: Robust direct and indirect impacts of education on the mortality rate

	Average treatment effect (ATE)				ATT	ATU
	direct effect		*indirect effect*		*direct effect*	
	$\theta(1)$	$\theta(0)$	$\eta(0)$	$\eta(1)$	$\theta(1)$	$\theta(0)$
Cox						
Primary to	-0.021	-0.072	-0.165^{+}	-0.115	0.020	-0.122^{+}
lower vocational	(0.067)	(0.048)	(0.077)	(0.062)	(0.069)	(0.051)
Lower vocational to	0.039	0.035	-0.092	-0.088	0.022	0.034
lower secondary	(0.035)	(0.038)	(0.047)	(0.050)	(0.037)	(0.040)
Lower secondary to	-0.127^{+}	-0.076	-0.064	-0.114	-0.090	-0.086
higher	(0.050)	(0.069)	(0.068)	(0.083)	(0.051)	(0.076)
Gamma-Gompertz						
Primary to	-0.019	-0.071	-0.179^{+}	-0.127^{+}	0.023	-0.121^{+}
lower vocational	(0.071)	(0.048)	(0.083)	(0.065)	(0.072)	(0.051)
Lower vocational to	0.042	0.043	-0.095	-0.096	0.025	0.042
lower secondary	(0.038)	(0.044)	(0.050)	(0.051)	(0.039)	(0.047)
Lower secondary to	-0.151^{+}	-0.094	-0.070	-0.128	-0.105	-0.109
higher	(0.061)	(0.081)	(0.081)	(0.098)	(0.062)	(0.090)

$^{+}p < 0.05$ and $^{**}p < 0.01$

Table B.8: Sensitivity analysis (mediator): outcome, selection and mediator effects (based on p^{δ}_{mj} in (17))

	Primary to lower vocational			Lower vocational to lower secondary			Lower secondary to higher		
	ω	ξ	ψ	ω	ξ	ψ	ω	ξ	ψ
Professional	-0.012	0.388^{**}	-0.417^{**}	0.010	0.708^{**}	-0.551^{**}	-0.054	0.182^{**}	-0.397^{**}
	(0.154)	(0.083)	(0.062)	(0.088)	(0.049)	(0.044)	(0.079)	(0.053)	(0.045)
Self-employed	0.154	0.169	-0.119^{**}	0.102	0.046	0.069	0.047	-0.191	0.256^{**}
	(0.169)	(0.106)	(0.075)	(0.109)	(0.069)	(0.062)	(0.118)	(0.098)	(0.075)
Skilled	-0.025	-0.096	0.195^{**}	-0.073	-0.315^{**}	0.336^{**}	0.013	-0.619^{**}	0.460^{**}
	(0.088)	(0.054)	(0.040)	(0.058)	(0.039)	(0.034)	(0.075)	(0.059)	(0.044)
Unskilled	0.017	-0.260^{**}	0.446^{**}	-0.012	-0.397^{**}	0.551^{**}	0.023	-0.636^{**}	0.642^{**}
	(0.109)	(0.064)	(0.049)	(0.080)	(0.056)	(0.047)	(0.097)	(0.097)	(0.064)
Missing	0.131	-0.155	0.311^{**}	0.241^{+}	-0.160^{+}	0.305^{**}	0.156	-0.227^{+}	0.293^{**}
	(0.153)	(0.098)	(0.078)	(0.111)	(0.077)	(0.071)	(0.133)	(0.119)	(0.090)
Born in Utrecht	0.286^{+}	-0.168	0.256^{**}	0.173	0.064	0.125^{+}	0.075	-0.077	0.106
	(0.133)	(0.087)	(0.071)	(0.100)	(0.066)	(0.062)	(0.108)	(0.088)	(0.073)
Catholic	-0.024	-0.058	0.136^{**}	-0.068	-0.111^{**}	0.173^{**}	0.005	-0.070	0.128^{**}
	(0.087)	(0.055)	(0.041)	(0.059)	(0.039)	(0.033)	(0.069)	(0.050)	(0.040)
Dutch Reformed	-0.044	0.113^{+}	-0.024	0.052	0.008	0.018	0.057	-0.020	0.020
	(0.095)	(0.056)	(0.042)	(0.061)	(0.040)	(0.033)	(0.070)	(0.050)	(0.040)
Calvinist	-0.009	0.400^{**}	-0.425^{**}	-0.001	0.119	-0.319^{**}	-0.135	0.049	-0.228^{**}
	(0.196)	(0.104)	(0.074)	(0.109)	(0.066)	(0.061)	(0.118)	(0.083)	(0.070)
Other religion	-0.286	0.260	0.061	-0.052	0.306	0.300	-0.262	-0.246	0.494^{+}
	(0.706)	(0.326)	(0.224)	(0.379)	(0.207)	(0.177)	(0.352)	(0.273)	(0.224)
Bad general health	0.080	-0.056	0.257^{**}	0.154^{+}	0.098^{+}	0.229^{**}	0.225^{+}	-0.063	0.329^{**}
	(0.112)	(0.062)	(0.053)	(0.072)	(0.049)	(0.045)	(0.079)	(0.061)	(0.049)
Bad hearing	0.395^{+}	-0.314^{+}	0.887^{**}	0.276	-0.225^{+}	0.942^{**}	0.228	-0.292	1.023^{**}
	(0.188)	(0.131)	(0.113)	(0.147)	(0.113)	(0.110)	(0.189)	(0.176)	(0.149)
Bad sight	0.040	0.448^{**}	-0.559^{**}	0.018	0.489^{**}	-0.565^{**}	0.033	0.192^{**}	-0.355^{**}
	(0.110)	(0.058)	(0.045)	(0.064)	(0.039)	(0.034)	(0.067)	(0.047)	(0.038)
Bad psychological	0.157	-0.476^{**}	0.824^{**}	0.150^{+}	-0.097^{+}	0.557^{**}	0.123	-0.187^{**}	0.507^{**}
	(0.091)	(0.053)	(0.048)	(0.068)	(0.046)	(0.042)	(0.079)	(0.064)	(0.050)

Based on adding U to propensity score with probabilities of U from observed probabilities for each covariate. No effect would give $\omega = 0, \xi = 0$ and $\psi = 0$. $^{+}p < 0.05$ and $^{**}p < 0.01$

Table B.9: Sensitivity analysis (mediator): direct effect IPW (ATE), based on p^{δ}_{mj} in (17).

	Primary to lower vocational		Lower vocational to lower secondary		Lower secondary to higher	
	$\theta(1)$	$\theta(0)$	$\theta(1)$	$\theta(0)$	$\theta(1)$	$\theta(0)$
neutral	−0.022	−0.070	0.038	0.036	−0.124	−0.080
	(0.069)	(0.048)	(0.036)	(0.038)	(0.050)	(0.069)
Professional	−0.022	−0.069	0.039	0.034	−0.124	−0.080
	(0.069)	(0.048)	(0.036)	(0.038)	(0.050)	(0.069)
Self-employed	−0.023	−0.071	0.037	0.036	−0.124	−0.079
	(0.068)	(0.048)	(0.035)	(0.038)	(0.050)	(0.069)
Skilled	−0.021	−0.069	0.037	0.033	−0.111	−0.072
	(0.068)	(0.048)	(0.036)	(0.038)	(0.052)	(0.072)
Unskilled	−0.022	−0.069	0.038	0.035	−0.118	−0.078
	(0.068)	(0.048)	(0.035)	(0.038)	(0.051)	(0.070)
Missing	−0.022	−0.070	0.038	0.037	−0.123	−0.079
	(0.068)	(0.048)	(0.035)	(0.038)	(0.050)	(0.069)
Born in Utrecht	−0.022	−0.069	0.037	0.034	−0.124	−0.079
	(0.068)	(0.048)	(0.035)	(0.038)	(0.050)	(0.069)
Catholic	−0.022	−0.069	0.038	0.035	−0.124	−0.080
	(0.068)	(0.048)	(0.035)	(0.038)	(0.050)	(0.069)
Dutch Reformed	−0.023	−0.069	0.038	0.036	−0.124	−0.079
	(0.068)	(0.048)	(0.035)	(0.038)	(0.050)	(0.069)
Calvinist	−0.022	−0.069	0.038	0.036	−0.125	−0.080
	(0.068)	(0.048)	(0.035)	(0.038)	(0.050)	(0.069)
Other religion	−0.022	−0.069	0.038	0.037	−0.125	−0.079
	(0.068)	(0.048)	(0.035)	(0.038)	(0.050)	(0.069)
Bad general health	−0.023	−0.070	0.033	0.032	−0.124	−0.079
	(0.068)	(0.048)	(0.036)	(0.038)	(0.050)	(0.069)
Bad hearing	−0.022	−0.069	0.038	0.036	−0.124	−0.078
	(0.068)	(0.048)	(0.035)	(0.038)	(0.050)	(0.069)
Bad sight	−0.022	−0.072	0.034	0.034	−0.124	−0.080
	(0.068)	(0.049)	(0.036)	(0.038)	(0.051)	(0.069)
Bad psychological	−0.023	−0.067	0.034	0.035	−0.124	−0.079
	(0.068)	(0.049)	(0.036)	(0.038)	(0.051)	(0.069)

Based on adding U to propensity score with probabilities of U from observed probabilities for observed variables.
Original estimates are $\theta(1)$: -0.022 (Primary to lower voc.); 0.038 (Lower voc. to lower sec.); -0.124 (Lower sec. to higher); $\theta(0)$: -0.070 (Primary to lower voc.); 0.036 (Lower voc. to lower sec.); -0.079 (Lower sec. to higher).

Table B.10: Sensitivity analysis (mediator): outcome, selection and mediator effects (based on p^M_{mi} in (18))

	Primary to lower vocational			Lower vocational to lower secondary			Lower secondary to higher		
	ω	ξ	ψ	ω	ξ	ψ	ω	ξ	ψ
Professional	0.022	0.149	-0.348^{**}	-0.004	0.577^{**}	-0.381^{**}	0.007	1.100^{**}	-0.366^{**}
	(0.151)	(0.085)	(0.067)	(0.096)	(0.055)	(0.048)	(0.086)	(0.054)	(0.044)
Self-employed	0.020	-0.047	-0.089	-0.004	0.216^{**}	0.021	0.003	-0.420^{**}	0.267^{**}
	(0.171)	(0.103)	(0.077)	(0.119)	(0.071)	(0.063)	(0.116)	(0.104)	(0.074)
Skilled	-0.002	-0.194^{**}	0.192^{**}	-0.005	-0.495^{**}	0.288^{**}	0.010	-1.102^{**}	0.352^{**}
	(0.087)	(0.054)	(0.040)	(0.058)	(0.041)	(0.035)	(0.076)	(0.075)	(0.050)
Unskilled	0.016	-0.503^{**}	0.445^{**}	-0.010	-0.498^{**}	0.464^{**}	-0.002	-1.048^{**}	0.589^{**}
	(0.104)	(0.061)	(0.049)	(0.079)	(0.059)	(0.049)	(0.104)	(0.120)	(0.072)
Missing	0.021	-0.221^{+}	0.315^{**}	-0.008	-0.196^{+}	0.278^{**}	-0.006	-0.310^{+}	0.258^{**}
	(0.157)	(0.096)	(0.076)	(0.125)	(0.081)	(0.071)	(0.145)	(0.127)	(0.089)
Born in Utrecht	0.019	-0.404^{**}	0.320^{**}	-0.008	0.105	0.144^{+}	0.001	0.131	0.108
	(0.141)	(0.087)	(0.069)	(0.115)	(0.068)	(0.062)	(0.116)	(0.087)	(0.072)
Catholic	-0.002	-0.336^{**}	0.138^{**}	-0.004	-0.098^{+}	0.102^{**}	0.012	0.047	0.070^{**}
	(0.088)	(0.054)	(0.040)	(0.059)	(0.041)	(0.033)	(0.071)	(0.050)	(0.039)
Dutch Reformed	0.011	0.291^{**}	-0.048	-0.004	0.018	0.014	0.012	-0.059	0.009
	(0.099)	(0.056)	(0.042)	(0.062)	(0.040)	(0.033)	(0.071)	(0.051)	(0.040)
Calvinist	-0.006	0.769^{**}	-0.493^{**}	-0.006	0.116	-0.305^{**}	-0.003	0.063	-0.211^{**}
	(0.234)	(0.121)	(0.079)	(0.111)	(0.068)	(0.062)	(0.116)	(0.084)	(0.071)
Other religion	-0.122	-0.024	0.107	-0.034	0.359	0.209	-0.014	0.262	0.477^{+}
	(0.646)	(0.345)	(0.254)	(0.444)	(0.228)	(0.198)	(0.383)	(0.261)	(0.216)
Bad general health	0.013	-0.223^{**}	0.301^{**}	-0.006	0.179^{**}	0.203^{**}	0.009	-0.030	0.334^{**}
	(0.112)	(0.061)	(0.054)	(0.080)	(0.049)	(0.045)	(0.083)	(0.061)	(0.050)
Bad hearing	0.018	-0.422^{**}	0.933^{**}	-0.017	-0.320^{+}	0.948^{**}	-0.013	-0.193	0.836^{**}
	(0.204)	(0.132)	(0.113)	(0.170)	(0.120)	(0.119)	(0.228)	(0.188)	(0.160)
Bad sight	0.017	0.207^{**}	-0.437^{**}	-0.005	0.604^{**}	-0.519^{**}	0.010	0.422^{**}	-0.347^{**}
	(0.109)	(0.059)	(0.048)	(0.068)	(0.041)	(0.035)	(0.067)	(0.047)	(0.038)
Bad psychological	0.009	-0.711^{**}	0.868^{**}	-0.006	-0.027	0.526^{**}	0.010	-0.092	0.499^{**}
	(0.091)	(0.053)	(0.047)	(0.071)	(0.047)	(0.041)	(0.082)	(0.063)	(0.049)

Based on adding U to propensity score with probabilities of U from observed probabilities for each covariate. No effect would give $\omega = 0, \xi = 0$ and $\psi = 0$. $^{+}p < 0.05$ and $^{**}p < 0.01$

Table B.11: Sensitivity analysis (mediator): direct effect IPW (ATE), based on p_{mi}^{M} in (18).

	Primary to lower vocational		Lower vocational to lower secondary		Lower secondary to higher	
	$\theta(1)$	$\theta(0)$	$\theta(1)$	$\theta(0)$	$\theta(1)$	$\theta(0)$
neutral	−0.022	−0.070	0.038	0.036	−0.124	−0.080
	(0.069)	(0.048)	(0.036)	(0.038)	(0.050)	(0.069)
Professional	−0.022	−0.070	0.037	0.036	−0.126	−0.083
	(0.068)	(0.048)	(0.036)	(0.038)	(0.055)	(0.079)
Self-employed	−0.022	−0.070	0.038	0.036	−0.125	−0.080
	(0.069)	(0.048)	(0.036)	(0.038)	(0.051)	(0.069)
Skilled	−0.022	−0.070	0.038	0.037	−0.124	−0.080
	(0.068)	(0.048)	(0.036)	(0.038)	(0.056)	(0.089)
Unskilled	−0.020	−0.068	0.037	0.035	−0.123	−0.079
	(0.068)	(0.049)	(0.036)	(0.038)	(0.053)	(0.075)
Missing	−0.022	−0.069	0.038	0.036	−0.124	−0.080
	(0.068)	(0.048)	(0.035)	(0.038)	(0.050)	(0.069)
Born in Utrecht	−0.022	−0.069	0.038	0.036	−0.124	−0.080
	(0.068)	(0.048)	(0.035)	(0.038)	(0.050)	(0.069)
Catholic	−0.023	−0.070	0.038	0.036	−0.125	−0.080
	(0.068)	(0.049)	(0.035)	(0.038)	(0.050)	(0.069)
Dutch Reformed	−0.023	−0.070	0.035	0.036	−0.124	−0.079
	(0.069)	(0.049)	(0.035)	(0.038)	(0.050)	(0.069)
Calvinist	−0.022	−0.070	0.038	0.036	−0.124	−0.079
	(0.068)	(0.049)	(0.035)	(0.038)	(0.050)	(0.069)
Other religion	−0.022	−0.070	0.038	0.036	−0.124	−0.079
	(0.068)	(0.045)	(0.035)	(0.038)	(0.050)	(0.069)
Bad general health	−0.022	−0.068	0.038	0.036	−0.124	−0.080
	(0.068)	(0.048)	(0.036)	(0.038)	(0.050)	(0.069)
Bad hearing	−0.022	−0.069	0.038	0.036	−0.124	−0.079
	(0.068)	(0.048)	(0.035)	(0.038)	(0.050)	(0.069)
Bad sight	−0.020	−0.071	0.036	0.036	−0.124	−0.080
	(0.069)	(0.048)	(0.036)	(0.038)	(0.051)	(0.069)
Bad psychological	−0.022	−0.065	0.038	0.035	−0.124	−0.080
	(0.068)	(0.048)	(0.036)	(0.038)	(0.050)	(0.069)

Based on adding U to propensity score with probabilities of U from observed probabilities for observed variables.
Original estimates are $\theta(1)$: -0.022 (Primary to lower voc.); 0.038 (Lower voc. to lower sec.); -0.124 (Lower sec. to higher); $\theta(0)$: -0.070 (Primary to lower voc.); 0.036 (Lower voc. to lower sec.); -0.079 (Lower sec. to higher).