Psychological Medicine

cambridge.org/psm

Disentangling potential causal effects of educational duration on well-being, and mental and physical health outcomes

Margot P. van de Weijer[1,2,3,†], Perline A. Demange[1,2], Dirk H.M. Pelt[1,2], Meike Bartels[1,2,*] and Michel G. Nivard[1,2,*]

[1]Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands; [2]Amsterdam Public Health Research Institute, Amsterdam University Medical Centres, Amsterdam, The Netherlands and [3]Genetic Epidemiology, Department of Psychiatry, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, The Netherlands

Original Article

*These authors contributed equally to this work.

†Current affiliation.

Cite this article: van de Weijer MP, Demange PA, Pelt DHM, Bartels M, Nivard MG (2024). Disentangling potential causal effects of educational duration on well-being, and mental and physical health outcomes. *Psychological Medicine* **54**, 1403–1418. https://doi.org/10.1017/S003329172300329X

Received: 7 October 2022
Revised: 4 October 2023
Accepted: 13 October 2023
First published online: 15 November 2023

Keywords:
causality; education; health; Mendelian randomization; well-being; within-family

Corresponding author:
Margot P. van de Weijer;
Email: m.p.vandeweijer@amsterdamumc.nl

Abstract

Background. Extensive research has focused on the potential benefits of education on various mental and physical health outcomes. However, whether the associations reflect a causal effect is harder to establish.

Methods. To examine associations between educational duration and specific aspects of well-being, anxiety and mood disorders, and cardiovascular health in a sample of European Ancestry UK Biobank participants born in England and Wales, we apply four different causal inference methods (a natural policy experiment leveraging the minimum school-leaving age, a sibling-control design, Mendelian randomization [MR], and within-family MR), and assess if the methods converge on the same conclusion.

Results. A comparison of results across the four methods reveals that associations between educational duration and these outcomes appears predominantly to be the result of confounding or bias rather than a true causal effect of education on well-being and health outcomes. Although we do consistently find no associations between educational duration and happiness, family satisfaction, work satisfaction, meaning in life, anxiety, and bipolar disorder, we do not find consistent significant associations across all methods for the other phenotypes (health satisfaction, depression, financial satisfaction, friendship satisfaction, neuroticism, and cardiovascular outcomes).

Conclusions. We discuss inconsistencies in results across methods considering their respective limitations and biases, and additionally discuss the generalizability of our findings in light of the sample and phenotype limitations. Overall, this study strengthens the idea that triangulation across different methods is necessary to enhance our understanding of the causal consequences of educational duration.

Introduction

There is an extensive body of research examining associations between educational attainment (EA) and mental and physical health outcomes. Existing studies have pointed to EA (measured as years of education, age at leaving education, or diploma obtained) as a correlate of well-being (Bücker, Nuraydin, Simonsmeier, Schneider, & Luhmann, 2018), depression (Lorant et al., 2003), quality-adjusted life years (Furnée, Groot, & Van Den Brink, 2008), different cardiovascular outcomes (Khaing, Vallibhakara, Attia, McEvoy, & Thakkinstian, 2017), and a wide range of other diseases and disorders (Choi et al., 2011; Putrik et al., 2016; Telfair & Shelton, 2012). Often, EA is interpreted as a modifiable risk factor that might improve outcomes in these different domains, but confounding and reverse causation are difficult to rule out.

Correlational evidence provides us with a first indication of associations between education and (mental) health outcomes. For example, a meta-analysis by Bücker et al. suggests a small-to-medium positive correlation between academic achievement and subjective well-being (SWB) that was stable across different measures of academic achievement and SWB (Bücker et al., 2018). Similarly, a small but significant correlation has been found between academic achievement and subsequent depression through meta-analysis (Huang, 2015). In addition, lower education has been associated with a higher risk of different cardiovascular outcomes (Khaing et al., 2017), and lower self-reported health (Furnée et al., 2008).

Such meta-analytic studies offer the opportunity to evaluate and summarize the existing literature, which allows us to identify correlations worth exploring in more detail. However, it is difficult to establish whether these associations reflect causal associations or whether they might be caused by residual confounding (e.g. genetics, socioeconomic status) (Fewell,

Davey Smith, & Sterne, 2007; Sobel, 2000). While confounders can be considered in meta-analysis, it is rarely the case that a large number of studies include the same confounders. Moreover, even if confounding factors could be ruled out, correlational studies would not offer clarity on the direction of causation. For example, while higher levels of education might lead to better access to healthcare, less health problems, and higher health (van der Heide et al., 2013), the reverse could also be true: for example, people in good health might have better possibilities to focus on education and reach higher levels of education than those in poor health (Kawachi, Adler, & Dow, 2010).

A quasi-experimental design that has been applied widely in educational research is to consider compulsory schooling laws where the legal minimum school-leaving age is increased (Brunello, Fort, & Weber, 2009; Clark & Royer, 2013; Glymour & Manly, 2018; Lleras-Muney, 2002) as an exposure over which individuals can be reasonably assumed to have no control. The implementation of these laws serves as a natural experiment where people are quasi-randomly separated in two groups (before and after, or subject to or not subject to the policy change). Assuming that this policy change only directly impacts the number of years someone stays in education, and assuming that is unrelated to confounding factors, this policy change can be used to estimate the direct effect of educational duration on diverse outcomes. Using this design, researchers have found positive effects of educational duration on mental health (Chevalier & Feinstein, 2006; Graeber, 2017), cognitive abilities (Banks & Mazzonna, 2012), mortality (Davies, Dickson, Smith, Van Den Berg, & Windmeijer, 2018), income (Davies et al., 2018; Grenet, 2013), and cardiovascular health (Hamad, Nguyen, Bhattacharya, Glymour, & Rehkopf, 2019). Nevertheless, there is still considerable disagreement across different studies employing this design due to heterogeneity in study features such as the included instrument, the examined number of years around the reform, or the populations included (see Hamad, Elser, Tran, Rehkopf, & Goodman, 2018). Additionally, the policy shift only affects those that would otherwise have left school earlier, meaning that we study a *Local Average Treatment Effect* (LATE) in this context. This is important to keep in mind when interpreting results, since this limits the generalizability of findings to those not affected by the reform (Ichino & Winter-Ebmer, 1999). For the subgroup of individuals affected by the reform, we also assume monotonicity, i.e. there are no individuals for whom the reform decreases their educational duration.

Another quasi-experimental design controlling for several forms of confounding using observational data is the sibling-control design. Comparing outcomes of biological siblings brought up in the same family allows to control for shared environmental confounding (e.g. socioeconomic conditions during childhood), and for shared genetic predispositions. However, factors unique to one of the siblings but not the other and measurement error can still bias the results of sibling-control studies (Frisell, 2021). Additionally, even if we could control for all unshared confounders, the method would not help us determine the direction of causation. If we find that siblings who score higher on well-being also stay in school longer, this could be because well-being causally increases school-leaving age, but the reverse is as likely: school-leaving age might causally increase well-being.

In Mendelian randomization (MR), one or more genetic variant(s) robustly associated with a predictor variable are used as instrumental variables to examine a potentially causal association between a predictor and outcome. The approach relies on Mendel's laws of segregation and independent assortment, which assume that genetic variants are inherited randomly from one's parents and independent from other genetic variants. Assuming that (1) the genetic variants are robustly associated with the exposure, (2) there are no unmeasured confounders of the instrument–outcome association, and (3) the genetic variants are not associated with the outcome of interest other than via the exposure (no pleiotropy), the genetic variants for an exposure can be used as instruments to examine potential causality between the exposure and an outcome. For example, a genetic variant associated with educational duration that is also indirectly associated with higher well-being (through its association with educational duration) provides supportive evidence of a causal association from education on well-being. Multiple studies have used MR to examine causal links between EA and health-related traits, with suggestive evidence for causal influences on traits like alcohol consumption, physical activity, and cardiovascular outcomes (Davies, Dickson, Davey Smith, Windmeijer, & van den Berg, 2019b; Gill, Efstathiadou, Cawood, Tzoulaki, & Dehghan, 2019). Importantly, these associations are only valid if the three key assumptions mentioned above are met. Unfortunately, it is often difficult to evaluate if the assumption of no pleiotropy is met, as many, or even most, genetic variants exert pleiotropic effects. In addition, unmodeled assortative mating, dynastic effects, and population stratification can spuriously induce associations between the genetic variant(s) and outcomes (Brumpton et al., 2020).

A further development of MR is the application of this method in the context of within-family analysis (Brumpton et al., 2020). By performing genetic instrumental variable within sibling pairs, we directly control for the influences of assortative mating, population stratification (siblings share the same population background), and dynastic effects. First, since genetic variants inherited by siblings are random within a family, genotype differences between siblings will be independent of assortative mating. Second, since the effects of parental wealth and status on their offspring is likely similar across siblings, genetic differences between siblings will be independent of dynastic effects. Lastly, genetic differences between siblings are independent of population stratification. Using within-sibling MR, Brumpton et al. demonstrate that conventional non-family MR estimates for the association between taller height/lower body mass index (BMI) and increased EA were almost entirely attenuated in the context of within-family MR (Brumpton et al., 2020). Similarly, Davies et al. used a sibling sample to check if identified associations between EA and different health measures were due to dynastic effects or assortative mating (Davies et al., 2019a, 2019b). They found little evidence that the within-family results were different from bivariate two-sample MR, but also note a probable lack of power.

While within-family MR has important advantages over conventional MR, it is nevertheless still fallible to unmet assumptions (e.g. the presence of pleiotropy) and is also less powerful as it is applied only in siblings within a larger sample. For both the conventional and the within-family MR, we assume *monotonicity* (i.e. the genetic variants do not have opposite effects in subgroups of people) and interpret identified effects as LATE.

There are various methods for examining causality in observational data, but all rely on strict assumptions that often are difficult to meet or evaluate. A way in which we can reduce our reliance on these individual assumptions is by applying multiple methods and evaluate the consistency of results and potential discrepancies therein, in light of the biases that accompany each of

https://doi.org/10.1017/S003329172300329X Published online by Cambridge University Press

these methods. In a study where the effect of BMI on different outcomes was assessed, the authors used both MR (subject to family-level confounding) and non-genetic and genetic within-family analyses (subject to reverse causation) (Howe et al., 2020). By verifying that these methods converge upon the same conclusion, the authors increase the certainty that the results were not a by-product of their respective biases. In a similar fashion, Davies et al. examined potential causal effects of education on health, mortality, and income using both a design where they leverage the raising of school-leaving age (ROSLA) and MR, with both methods suggesting similar effects for almost all outcomes (Davies, Dickson, Davey Smith, Windmeijer, & van den Berg, 2021).

For the current project, we are interested in causal influences on specific aspects of well-being, anxiety and mood disorders, and cardiovascular health. As educational effects on well-being are of primary interest to us, we depart from treating 'well-being' as a single unified outcome and separately consider effects on satisfaction with family relations, work, friendships, health, and finances (Schimmack, 2008). We rely on four widely accepted techniques for causal inference: we make use of a random natural policy shift in England and Wales in September 1972 that raised school-leaving age from 15 to 16 but is unlikely to be related to confounding factors. We perform analyses within sibships to control for shared environmental confounders, and partly control for shared genetics. We make use of an index of genetic variation related to EA as an instrumental variable in MR. Finally, we combine the genetic instrumental variable with within-family analysis in sibling pairs. We apply those techniques in a single homogenously measured sample (the UK Biobank [UKB]), minimizing variation in results due to differences in measurement. By assessing if these different methods converge on the same conclusion in terms of whether or not there is a causal effect of educational duration on the different outcomes, we can be more confident in our conclusions on the potential causal relation between education and the different outcomes.

Methods

This project was pre-registered at the Open Science Framework (https://osf.io/s6gha). Deviations from the pre-registration are indicated throughout the manuscript.

Sample

We used data from the UKB, a large UK cohort study which collected genetic and phenotypic data on ±500 000 participants between 40 and 69 years old at recruitment (Bycroft et al., 2018). For the current project, we selected individuals of European ancestry (a decision taken to minimize ancestral confounding in genetic analyses) that were born in England and Wales (to ensure participants were likely affected by the school-leaving age reform). Specific further sample selection procedures for the four different analyses are described below per analysis, and a flowchart of sample selection per analysis is found in online Supplementary Fig. S1.

Education variable

We used UKB data-field 845 'age completed full-time education' as our education exposure variable. Participants were asked to answer the question 'at what age did you complete your continuous full-time education?'. If someone provided an answer below 5, or an answer higher than their age, the answer was rejected. If someone answered with an age higher than 40, the participant was asked to confirm their answer. Since the question was not collected in participants who indicated having a college or university degree, we, in line with the literature (Davies et al., 2018; Plotnikov et al., 2020), imputed their age at completed full-time education as 21. In case someone provided an answer on more than one instance, we used the last available answer as the age at which one completed their full-time education. If the answer at the later time-point indicated a lower age than a previous answer ($N = 72$), we coded the answer as missing.

Outcome variables

General information on item construction and cleaning procedures for these variables can be found in the Supplementary Methods. The following self-report items were included as well-being outcome variables: *general happiness* based on happiness (UKB ID 4526) and general happiness (UKB ID 20459), *family relationship satisfaction* (UKB ID 4559), *financial situation satisfaction* (UKB ID 4581), *friendship satisfaction* (UKB ID 4570), *work/job satisfaction* (UKB ID 4537), *health satisfaction* based on health satisfaction (UKB ID 4548) and general happiness with own health (UKB ID 20459), and *belief that own life is meaningful* (UKB ID 20460). All items were coded so that a higher score indicated a higher level of well-being. For *neuroticism*, we included a summary score (UKB ID 20127) that was based on 12 neurotic domain self-report items. We used a combination of medical record data (UKB ID 41270) and self-report data (UKB ID 20002) to create binary variables reflecting if someone was ever diagnosed with *depression*, *anxiety*, or *manic or bipolar disorder*. Lastly, a binary variable indicating *cardiovascular problems* was constructed based on vascular/heart problems diagnosed by a doctor (UKB ID 6150) or self-reported (UKB ID 20002).

Control outcomes

We selected four negative control outcomes: *height* (UKB ID 50), *birthweight* (UKB ID 20022), *comparative body size at age 10* (UKB ID 1687), and *comparative height size at age 10* (UKB ID 1697). It is unlikely these variables are causally influenced by additional years of schooling, but the presence of confounding parental variables (e.g. parental SES) might lead to observable but false-positive associations. As a positive control outcome, we included *average total household income before tax* (UKB ID 738), which was split into the four yes/no dichotomous variables: income over 18k, income over 31k, income over 52k, and income over 100k. General information on item construction and cleaning procedures for these variables can also be found in the Supplementary Methods.

Covariates

As phenotypic covariates, we included sex (UKB ID 31), assessment center (UKB ID 54), family size (based on number of [adopted] siblings, UKB IDs 1873, 3972, 1883, and 3982), season of birth (based on month of birth, UKB ID 52), and year of birth (UKB ID 34). Genetic covariates included the first 10 genomic principal components (PCs) and batch (UKB ID 22000).

https://doi.org/10.1017/S003329172300329X Published online by Cambridge University Press

Genotype data

Single-nucleotide polymorphisms (SNPs) from HapMap3 (CEU: Utah residents with Northern and Western European Ancestry) (1 345 801 SNPs) were filtered out of the imputed dataset. A pre-principal component analysis (PCA) quality control (QC) was done on unrelated individuals, filtering out SNPs with minor allele frequency (MAF) <0.01 and missingness >0.05, leaving 1 252 123 SNPs. After filtering out individuals with non-European ancestry, the SNP QC was repeated on unrelated Europeans ($N = 312\,927$). SNPs with MAF <0.01, missingness >0.05, and Hardy-Weinberg equilibrium (HWE) $p < 10^{-10}$ were filtered, leaving 1 246 531 SNPs. The HWE p-value threshold of 10^{-10} was based on: http://www.nealelab.is/blog/2019/9/17/genotyped-snps-in-uk-biobank-failing-hardy-weinberg-equilibrium-test. A final dataset of 1 246 531 QC-ed SNPs was created for 456 028 UKB subjects of European ancestry.

UKB correction

While the UKB is a valuable dataset where a large number of participants have been genotyped and extensively phenotyped, it is not necessarily representative of the UK population due to confounding from volunteer bias (Batty, Gale, Kivimäki, Deary, & Bell, 2020). To partially correct for volunteer bias, we calculate and include inverse probability weights using procedures by van Alten, Domingue, Galama, and Marees (2022). The respondents are weighted using weights based on sex, year of birth (5-year cohort), education level, ethnicity, region of residence (Census Greater London Area), tenure of dwelling, employment status, number of cars in the household, a dummy indicating whether the person lives in a single-person household, and self-reported health. For a more detailed description, see van Alten et al. (2022).

Analyses

We use four different methods to examine potential causal effects between educational duration and our outcomes. Table 1 provides an overview of these four methods, including their respective advantages and limitations. Sample descriptives per method can be found in Table 2. Below, we describe each of the four methods in more detail. All analysis code is available at https://github.com/margotvandeweijer/EA_causality. All continuous outcomes were standardized so that the resulting effect sizes reflect the S.D. increase in the outcomes for each additional year of education (see Table 2 for an overview of the S.D.s of the included variables).

Instrumental variable analysis leveraging the ROSLA

We used the ROSLA policy reform where the minimum school-leaving age was increased from 15 to 16 in England and Wales to examine the effects of longer schooling on our different outcomes. We selected a sample of UKB participants born in a 5-year window (1 February 1955 to 1 February 1960) around the reform (1 September 1972), and excluded related individuals (KING kinship coefficient >0.0884) using the *ukbtools* package in R (Hanscombe, Coleman, Traylor, & Lewis, 2019). A binary ROSLA indicator was created for this subset of participants that indicates if a participant was born before (affected = 0) or after (affected = 1) 1 September 1957 and was thus affected by the reform or not. Additionally, we transformed the age at which one left full-time education variable into a binary variable that indicates if an individual stayed in school after age 15 or not (Davies et al., 2019a). Next, we used two-stage least squares (2SLS) instrumental variable analyses using the *fixest* R package (Bergé, 2018), where in the first stage the binary education variable was included as the dependent variable and the binary ROSLA indicator was included as the instrument. In the second stage, we regressed all our standardized outcome variables on the fitted education values from the first-stage regression. Both stages included the phenotypic covariates. For comparative purposes, we also run regular (non-pre-registered) ordinary least squares (OLS) regression in the same sample the binary education predictor was used to predict the different outcomes (including the same covariates as the ROSLA analyses). To examine the robustness of the ROSLA results, we repeated the analyses using samples born in a 2 and 10 years window around the reform.

Sibling control design

We perform analyses within sibships to control for shared familial background characteristics, and partly control for genetic effects. Biological sibships in the UKB dataset are defined as participants with a kinship coefficient between $\frac{1}{2^{5/2}}$ and $\frac{1}{2^{3/2}}$ and a probability of zero identical-by-state sharing >0.0012 (Bycroft et al., 2018; Manichaikul et al., 2010). Individuals indicating they were adopted were removed from this sample. For each sibship j with i siblings, we start by calculating the average age at which sibships left full-time education $\overline{edu_{oj}} = \sum_1^m edu_{ij}/m$. Next, we calculate each sibling's deviation from the sibship average: $edu_{\Delta ij} = edu_{ij} - \overline{edu_{0j}}$. We use these estimates in a linear model where each outcome Y_{ij} for sibling i in sibship j is predicted as follows:

$$Y_{ij} = \beta_{00} + \beta_B \overline{edu_{oj}} + \beta_W edu_{\Delta ij} + covariates + e$$

where β_B is the between-sibship effect estimating if the average school-leaving age within sibships is associated with our outcomes, and β_W is the within-sibship effect estimating if a sibling deviating from the sibship school-leaving age average is associated with our outcome measures. Since we examine the effect of these within- and between-sibship estimates on the outcomes of individual siblings, we excluded sibships where only one sibling reported on educational duration, but we did not exclude sibships where not all siblings reported on one or more outcome measures. We report robust standard errors taking into account familial clustering, calculated using the coeftest function from the *lmtest* r-package (Hothorn et al., 2022). All phenotypic covariates were included in the analyses.

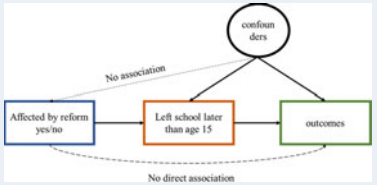
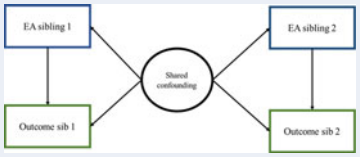
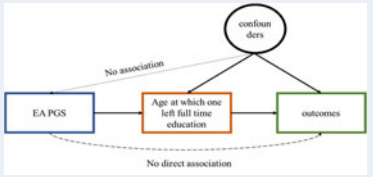
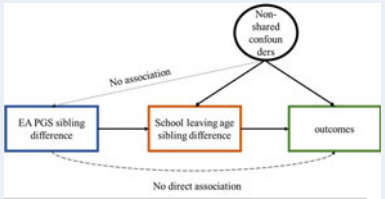
Mendelian randomization

We used polygenic scores (PGS) for EA in 2SLS instrumental variable analysis as genetic instruments for testing a directed causal association between educational duration and the outcomes. PGS are aggregate measures of genetic susceptibility for a trait of interest weighted by effect size estimates from genome-wide association studies (Choi, Mak, & O'Reilly, 2020). To calculate the PGS for EA, we used the summary statistics from the Genome Wide Association Study (GWAS) of years of education by Lee et al. (2018), excluding 23andme and British cohorts ($N = \sim$245k). PGS were constructed from the set of genome-wide significant HapMap3 SNPs ($p < 5 \times 10^{-8}$), pruned to be independent (using the package *TwoSampleMR* [Hemani et al., 2018]) using a clumping window of 1000 kb and a linkage disequilibrium (LD) cut-off of $R^2 = 0.1$. The PGS prediction accuracy for EA was assessed based on the incremental R^2 when including the PGS in a regression with all covariates.

Next, the PGS was used as a genetic instrument in 2SLS instrumental variable analysis in a sample of unrelated UKB

https://doi.org/10.1017/S003329172300329X Published online by Cambridge University Press

Table 1. Overview of different methods used in the present study

Method	Short summary	Core assumptions	Core limitations*	Visual description
ROSLA reform IV analysis	On 1 September 1972, the raising of school-leaving age in England and Wales was raised from 15 to 16. As a result, the compulsory school stay for individuals born in September 1957 and later was a year longer than for those born before September 1957. We used this policy shift as an instrument in instrumental variable analysis, where the reform directly affects school-leaving age, but does not directly affect any of our outcomes	(1) Relevance: the ROSLA reform associates with the exposure, (2) independence: no unmeasured confounders of the instrument–outcome association, (3) exclusion: no uncontrolled effect of the instrument on the outcome except via the exposure, (4) monotonicity	The reform only directly impacts those who would have otherwise left school at age 15. Thus, the instrument only affects a small part of the population and results are not generalizable to the entire population	
Sibling control design	The chances and support provided by the (early) childhood (shared) environment is considered one of the primary causes of confounding in educational research, these can be controlled for by relating outcomes to differences in the educational measure within sibling pairs	There are no unmeasured confounders of the EA–outcome association conditional on a familial effect	Confounders unshared by family members and measurement error can still bias the results	
Mendelian randomization (MR)	MR is a special form of instrumental variable analysis, where the instrument is based on genetic variants associated with the exposure. By using genetic variants as instrumental variants, MR is unlikely to be subject to reverse causality. We used a PGS for EA based on Lee et al. (2018) as the instrument in our analyses	(1) Relevance: the EA PGS is associated with EA, (2) independence: there are no unmeasured confounders of the PGS–outcome association, (3) exclusion: the PGS only affects the outcome via its effect on EA, (4) monotonicity	Key assumptions of the method, like no pleiotropy, do not always hold. Additionally, residual confounders such as dynastic effects can influence the results	
Mendelian randomization in sibships	This method is a combination of MR and the sibling control method; it has the same advantages as MR but additionally controls for assortative mating, dynastic effects, and population stratification. We perform MR in a sample of sibships where we take the difference between the sibships on the PGS and school-leaving age to remove the effect of family-level confounders	The same assumptions as MR and the sibling control design	Can still be confounded by unmet assumptions, and is less powerful than conventional MR as the sample is reduced to only sibships	

*All methods are susceptible for bias from selection/collider bias.

Table 2. Sample descriptives full sample, and per analysis type (for those with education data)

Full sample				ROSLA			Sibling control			MR		
Education												
	M **(s.d.)**	**Range**	***N*** **(females/males)***	***M*** **(s.d.)**	**Range**	***N*** **(females/males)***	***M*** **(s.d.)**	**Range**	***N*** **(females/males)***	***M*** **(s.d.)**	**Range**	***N*** **(females/males)***
Age when left full-time education	17.93 (2.71)	5–35	361 945 (196 936/168 613)	18.34 (2.57)	5–35	47 667 (26 620/20 966)	17.77 (2.70)	5–35	31 337 (18 068/13 269)	17.95 (2.72)	5–35	335 076 (179 164/155 912)
Continuous outcomes												
	M **(s.d.)**	**Range**	***N*** **(females/males)***	***M*** **(s.d.)**	**Range**	***N*** **(females/males)**	***M*** **(s.d.)**	**Range**	***N*** **(females/males)**	***M*** **(s.d.)**	**Range**	***N*** **(females/males)**
Happiness	4.5 (0.73)	1–6	208 936 (113 078/95 481)	4.42 (0.75)	1–6	28 418 (16 334/12 032)	4.50 (0.72)	1–6	17 595 (10 248/7347)	4.50 (0.73)	1–6	192 465 (103 629/88 836)
Health satisfaction	4.29 (0.89)	1–6	209 132 (113 191/95 563)	4.27 (0.92)	1–6	28 472 (16 368/12 052)	4.31 (0.88)	1–6	17 620 (10 266/7354)	4.29 (0.89)	1–6	192 643 (103 739/88 904)
Family satisfaction	4.8 (0.89)	1–6	152 202 (81 118/70 799)	4.72 (0.93)	1–6	20 162 (11 487/8878)	4.82 (0.86)	1–6	12 562 (7195/5367)	4.80 (0.89)	1–6	140 133 (74 294/65 839)
Financial satisfaction	4.36 (0.94)	1–6	152 882 (81 330/71 265)	4.23 (1.01)	1–6	20 246 (11 381/8826)	4.38 (0.91)	1–6	12 549 (7187/5362)	4.36 (0.94)	1–6	140 843 (74 533/66 310)
Friendship satisfaction	4.77 (0.74)	1–6	151 866 (81 028/70 554)	4.70 (0.77)	1–6	20 076 (11 318/8720)	4.77 (0.73)	1–6	12 458 (7160/5298)	4.77 (0.74)	1–6	139 869 (74 238/65 631)
Work satisfaction	4.41 (0.87)	1–6	100 993 (52 949/47 844)	4.28 (0.90)	1–6	17 499 (9762/7700)	4.40 (0.87)	1–6	8253 (4641/3612)	4.41 (0.87)	1–6	92 871 (48 427/44 444)
Meaning in life	3.7 (0.83)	1–5	115 434 (64 848/50 397)	3.66 (0.87)	1–5	16 890 (10 241/6623)	3.72 (0.82)	1–5	10 119 (6075/4044)	3.69 (0.82)	1–5	106 741 (59 640/47 101)
Neuroticism	4.11 (3.25)	0–12	297 342 (157 409/139 407)	4.42(3.33)	0–12	39 498 (21 891/17 547)	4.08 (3.24)	0–12	25 478 (14 483/10 995)	4.10 (3.25)	0–12	272 531 (143 344/129 187)
Binary outcomes												
		N **diagnosed**			***N*** **diagnosed**			***N*** **diagnosed**			***N*** **diagnosed**	
Depression		31 577			4707			2621			28 643	
Anxiety		13 799			1795			1142			12 458	
Bipolar or manic disorder		1516			236			136			1368	
Cardiovascular problems		114 926			10 414			9706			105 166	

Control outcomes												
Continuous	***M* (s.d.)**	**Range**	***N***	***M* (s.d.)**	**Range**	***N***	***M* (s.d.)**	**Range**	***N***	***M* (s.d.)**	**Range**	***N***
Birthweight	3.39 (0.42)	2.5–4.5	180 536 (108 979/71 235)	3.39 (0.42)	2.5–4.5	27 908 (17 023/10 829)	3.39 (0.42)	2.5–4.5	15 543 (9898/5645)	3.39 (0.42)	2.5–4.5	165 453 (99 463/65 990)
Height	168.8 (9.23)	138–202.5	364 701 (196 936/168 613)	169.5 (9.16)	138–202.5	47 483 (26 568/20 915)	168.1 (9.15)	138–202.0	31 280 (18 045/13 235)	168.9 (9.23)	138–202.5	334 307 (178 786/155 521)
Categorical	***N* below average**	***N* average**	***N* above average**	***N* below average**	***N* average**	***N* above average**	***N* below average**	***N* average**	***N* above average**	***N* below average**	***N* average**	***N* above average**
Comparative body size at age 10	116 778	181 760	56 596	14 987	23 123	8287	10 291	15 670	4471	106 417	166 606	52 039
Comparative height size at age 10	72 273	193 564	90 747	9492	25 062	11 994	323	16 686	7221	65 746	177 018	83 692

*For some participants, sex information is missing.

participants (KING kinship coefficient >0.0884). In the first stage, we predicted age at which one left full-time education (standardized) from the PGSs. In the second stage, the outcome and control outcomes were predicted from the fitted education values. All phenotypic and genetic covariates were included as covariates in both stages. The MR analyses were conducted using the *fixest* package in R (Bergé, 2018). For comparison, we also perform regular (non-pre-registered) OLS regression in the same sample, where standardized age at which one left full-time education is used to predict the outcomes, whilst correcting for the phenotypic covariates.

Mendelian randomization in sibships

Since one of the limitations of MR is its susceptibility to residual confounding stemming from dynastic effects, population stratification, and assortative mating, we additionally perform MR within sibships. We identify siblings in UKB and calculate each sibling's deviation from the sibship average using the same methodology as used for the sibling control design (see 'Sibling control design'). Additionally, we use the PGSs calculated for the MR analyses (see 'Mendelian randomization') to calculate a PGS average within sibships: $\overline{PGS_{oj}} = \sum_1^m PGS_{ij}/m$, and each sibling's deviation from the sibship average: $PGI_{\Delta ij} = PGI_{ij} - \overline{PGI_{0j}}$. We use these deviation estimates in instrumental variable regression (using the *fixest* package), where in the first stage we predict the sibling education deviation from the sibling PGS deviation. Next, the outcome and control outcomes were predicted from the first-stage fitted education values. Similar to the within-sibling analyses, we excluded sibships where only one sibling reported on EA, but we did not exclude sibships where not all siblings reported on one or more outcome measures. We report robust standard errors taking into account familial clustering, calculated using the coeftest function from the *lmtest* r-package (Hothorn et al., 2022). Both the phenotypic and genetic covariates were included.

Pre-registered interpretation of results

We define an unambiguous causal association as one where the policy shift, the sibling control design, and the Mendelian randomization analyses all imply a significant result in the same direction. The absence of significance across these methods would imply the absence of such a result. Due to the lower power associated with our within-sibship MR analyses, we are satisfied if the magnitude and direction of the Mendelian randomization within siblings is consistent with the other methods. With respect to statistical significance and multiple testing, we use two significance thresholds: (1) a suggestive threshold where we correct for the number of outcomes (15), so that $\alpha = 0.05/15 = 0.003$, and (2) a conservative threshold where we correct for the number of outcomes (15) and analysis types (4), so that $\alpha = 0.05/60 = 0.0008$. Inconsistencies across results will be interpreted along the potential biases and assumptions that accompany the different methods.

Results

Instrumental variable analysis leveraging the ROSLA

Table 3 depicts the results of the ROSLA instrumental variable analyses. Based on the 2SLS models, none of the outcomes are significantly predicted by age at which one left full-time education. This contrasts our comparative OLS analyses, which do not

https://doi.org/10.1017/S003329172300329X Published online by Cambridge University Press

Table 3. Results ROSLA instrumental variable analyses

	Education (fitted)			*F*-test (1st stage)		Wu–Hausman[a]		Regular OLS education	
Main outcomes	β (S.E.)	*p*	*N*	*F*	*p*	*wh*	*p*	β (S.E.)	*p*
Happiness	0.11 (0.15)	0.525	27 434	583.2	$<2.2 \times 10^{-16}$	1.79	0.181	−0.03 (0.01)	0.009
Health satisfaction	0.09 (0.15)	0.525	27 434	583.2	$<2.2 \times 10^{-16}$	1.79	0.181	0.02 (0.01)	0.072
Family satisfaction	0.15 (0.16)	0.346	19 392	457.2	$<2.2 \times 10^{-16}$	0.640	0.424	0.02 (0.02)	0.283
Financial satisfaction	−0.24 (0.17)	0.160	19 474	444.0	$<2.2 \times 10^{-16}$	11.5	0.0007	0.33 (0.03)	**$<2.2 \times 10^{-16}$**
Friendship satisfaction	0.04 (0.16)	0.814	19 310	458.2	$<2.2 \times 10^{-16}$	0.305	0.581	−0.05 (0.02)	0.060
Work satisfaction	−0.07 (0.18)	0.697	16 895	369.9	$<2.2 \times 10^{-16}$	0.519	0.471	0.06 (0.03)	0.021
Meaning in life	0.43 (0.29)	0.130	16 341	216.9	$<2.2 \times 10^{-16}$	2.03	0.154	0.03 (0.03)	0.340
Neuroticism	0.05 (0.10)	0.630	38 187	958.2	$<2.2 \times 10^{-16}$	8.44	0.004	−0.25 (0.02)	**$<2.2 \times 10^{-16}$**
Depression	−0.04 (0.03)	0.180	45 840	1205.9	$<2.2 \times 10^{-16}$	1.25	0.264	−0.07 (0.004)	**$<2.2 \times 10^{-16}$**
Anxiety	0.01 (0.02)	0.529	45 840	1205.9	$<2.2 \times 10^{-16}$	4.97	0.026	−0.03 (0.003)	**$<2.2 \times 10^{-16}$**
Bipolar or manic disorder	−0.003 (0.007)	0.714	45 840	1205.9	$<2.2 \times 10^{-16}$	0.373	0.542	−0.007 (0.001)	**9.42×10^{-10}**
Cardiovascular problems	0.02 (0.04)	0.673	45 840	1205.9	$<2.2 \times 10^{-16}$	5.73	0.017	−0.07 (0.006)	**$<2.2 \times 10^{-16}$**
Control outcomes									
Income over 18k	0.02 (0.04)	0.647	42 079	1088.7	$<2.2 \times 10^{-16}$	28.0	1.22×10^{-7}	0.20 (0.006)	**$<2.2 \times 10^{-16}$**
Income over 31k	0.14 (0.04)	0.002	42 079	1088.7	$<2.2 \times 10^{-16}$	9.49	0.002	0.27 (0.007)	**$<2.2 \times 10^{-16}$**
Income over 52k	0.10 (0.04)	0.016	42 079	1088.7	$<2.2 \times 10^{-16}$	5.69	0.017	0.21 (0.007)	**$<2.2 \times 10^{-16}$**
Income over 100k	0.02 (0.02)	0.432	42 079	1088.7	$<2.2 \times 10^{-16}$	2.43	0.119	0.05 (0.004)	**$<2.2 \times 10^{-16}$**
Birthweight	0.003 (0.12)	0.977	26 921	736.1	$<2.2 \times 10^{-16}$	0.045	0.831	0.02 (0.02)	0.247
Height	−0.04 (0.06)	0.479	45 743	1194.8	$<2.2 \times 10^{-16}$	11.2	0.0008	0.16 (0.010)	**$<2.2 \times 10^{-16}$**
Comparative body size at age 10	−0.01 (0.06)	0.813	44 650	1182.5	$<2.2 \times 10^{-16}$	0.028	0.867	−0.003 (0.01)	0.756
Comparative height size at age 10	−0.16 (0.06)	0.008	44 795	1192.7	$<2.2 \times 10^{-16}$	12.2	0.0005	0.06 (0.01)	**3.37×10^{-9}**

Note. All **continuous** outcomes were standardized. Assessment center, sex, season of birth, and year of birth were included as covariates.
p-values indicated in bold are lower than the conservative *p*-value threshold of 0.0008.
[a]H0 is the absence of endogeneity of the instrumented variables.

control for unmeasured confounders, where most associations were significant. The *F*-statistic of the 2SLS analyses ranged from 216.9 to 1205.9 depending on the outcome of interest, indicating that our instrument is unlikely to suffer from weak instrument bias. Since the standard errors are relatively large and the Wu–Hausman statistics, which test for the absence of endogeneity, were almost always non-significant at $\alpha = 0.05$, it is suggested that the 2SLS and OLS models do not statistically differ. However, the methods do lead to different estimates, suggesting the OLS results are nonetheless subject to considerable bias. Examining these associations in a 2- or 10-year window around the reform did not change our conclusions (see online Supplementary Table S1).

These findings contrast earlier findings by Davies et al. (2018). Using instrumental variable regression in UKB, they did observe an effect of remaining in school after age 15 on different cardiovascular outcomes and income. The main difference between the current study and the Davies et al. study is the method of correcting for year of birth, where they used a difference-in-difference approach instead of including this variable as a covariate. Therefore, we performed supplementary (non-preregistered) analyses where we, in a step-wise fashion, added season of birth and year of birth. The results are shown in online Supplementary Table S2 and Fig. 1. While adding year of birth as covariates might increase the chance that we are overcorrecting, it is evident from these results that the use of a policy experiment as an instrumental variable is very sensitive to the model specification: inclusion year of birth renders previously significant associations with happiness, familial, financial, and work satisfaction, cardiovascular problems, income, birthweight, and height non-significant.

Sibling control design

In total, there were 15 237 families with sibships of at least two siblings. The number of included individuals per outcome varied (see Table 4 for the sample size per outcome). The intra-class correlation for education, reflecting the amount of total variation in education explained by the family-level, was 0.40. Table 4 presents the within- and between-sibship estimates from the sibling control analyses. For the main outcomes, the between-sibling estimates for school-leaving age were significantly associated (based on the conservative $\alpha = 0.0008$ threshold) with happiness, health satisfaction, family satisfaction, financial satisfaction, work satisfaction, neuroticism, anxiety, and cardiovascular problems. However, the within-sibling estimates (indicating a potential causal effect) were only significant for financial satisfaction ($\beta = 0.025$, S.E. $= 0.006$, $p = 9.70 \times 10^{-6}$), and neuroticism ($\beta = -0.016$, S.E. $= 0.004$, $p = 3.67 \times 10^{-5}$), indicating a positive association between longer education and financial satisfaction and a negative association with neuroticism. With respect to the positive control outcomes, all between- and within-sibship estimates were significant. Between-sibling estimates for negative control outcomes also showed significant positive associations with age at leaving school, except for comparative body size at age 10. As expected, the within-sibling estimates were however not significant for birthweight and comparative height/body size at age 10, but surprisingly still significant for height, suggesting that the within-family estimates are not adequately correcting for all sources of bias.

Mendelian randomization

The EA PGS predicted 0.28% of the variance in school-leaving age, which is similar to the predictive power of the EA PGS

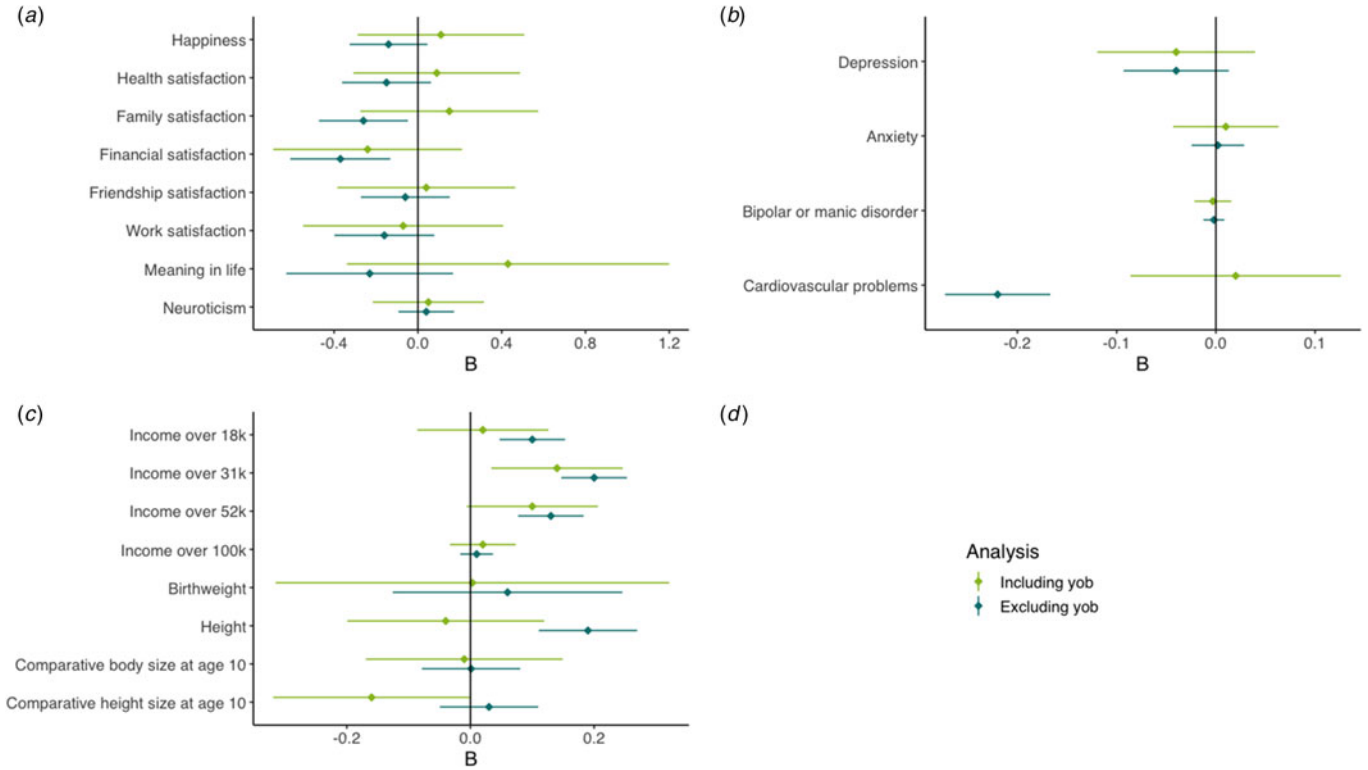


Figure 1. Comparison of ROSLA results including and excluding year of birth (yob) as a covariate for (a) continuous outcome measures, (b) binary outcome measures, and (c) control measures.

https://doi.org/10.1017/S003329172300329X Published online by Cambridge University Press

Table 4. Results sibling control analyses

		Within-sibling estimate			Between-sibling estimate		
Main outcomes	*N*	*β* (S.E.)	*t*	*p*	*β* (S.E.)	*t*	*p*
Happiness	17 071	−0.005 (0.005)	−0.941	0.347	−0.014 (0.004)	−3.989	**6.68×10^{-5}**
Health satisfaction	17 096	0.002 (0.005)	0.393	0.694	0.015 (0.004)	4.068	**4.76×10^{-5}**
Family satisfaction	12 137	−0.002 (0.006)	−0.363	0.717	−0.022 (0.004)	−5.203	**1.99×10^{-7}**
Financial satisfaction	12 132	0.025 (0.006)	4.426	**9.70×10^{-6}**	0.032 (0.004)	7.448	**1.01×10^{-13}**
Friendship satisfaction	12 042	−0.006 (0.006)	−1.07	0.284	−0.033 (0.004)	−7.680	**1.72×10^{-14}**
Work satisfaction	7979	−0.001 (0.007)	−0.176	0.860	0.001 (0.005)	0.266	0.790
Meaning in life	9885	0.010 (0.007)	1.451	0.147	−0.015 (0.005)	−3.117	0.002
Neuroticism	24 680	−0.016 (0.004)	−4.128	**3.67×10^{-5}**	−0.031 (0.003)	−9.997	**1.74×10^{-23}**
Depression	30 192	0.0003 (0.001)	0.332	0.740	−0.001 (0.0008)	−1.809	0.070
Anxiety	30 192	0.001 (0.0007)	1.659	0.097	−0.002 (0.0005)	−3.817	**0.0001**
Bipolar or manic disorder	30 192	0.00005 (0.0003)	0.178	0.859	0.0006 (0.0002)	3.298	0.001
Cardiovascular problems	30 192	−0.003 (0.002)	−1.796	0.073	−0.012 (0.001)	−9.132	**7.12×10^{-20}**
Control outcomes							
Income over 18k	26 678	0.020 (0.002)	12.079	**1.67×10^{-33}**	0.040 (0.001)	33.164	**2.32×10^{-236}**
Income over 31k	26 678	0.025 (0.002)	12.737	**4.74×10^{-37}**	0.052 (0.001)	36.853	**4.93×10^{-290}**
Income over 52k	26 678	0.021 (0.002)	12.851	**1.09×10^{-37}**	0.040 (0.001)	32.427	**2.82×10^{-226}**
Income over 100k	26 678	0.005 (0.001)	8.086	**6.42×10^{-16}**	0.011 (0.001)	17.884	**4.04×10^{-71}**
Height	30 410	0.012 (0.002)	5.936	**2.94×10^{-9}**	0.040 (0.002)	17.870	**4.73×10^{-71}**
Birthweight	15 052	0.006 (0.005)	1.268	0.205	0.020 (0.004)	4.959	**7.15×10^{-7}**
Comparative body size at age 10	29 341	0.003 (0.002)	1.291	0.197	0.005 (0.002)	2.359	0.018
Comparative height size at age 10	29 379	0.007 (0.002)	2.878	0.004	0.016 (0.002)	7.462	**8.76×10^{-14}**

p-values indicated in bold are lower than the conservative *p*-value threshold of 0.0008.

from Lee et al. that was based on the genome-wide significant SNPs based on a similarly sized ($N = 293\,723$) discovery GWAS (Okbay et al., 2016). Despite the relatively low predictive power, the *F*-values from our MR analyses ranged from 240.1 to 957.1 for the different outcomes, indicating that the PGS did not suffer from weak instrument bias.

The full results from the MR analyses are shown in Table 5. Six outcomes were significantly associated with school-leaving age based on our conservative significance threshold of $\alpha = 0.0008$: we found positive associations with health satisfaction ($\beta = 0.37$, S.E. = 0.05, $p = 1.31 \times 10^{-14}$) and financial satisfaction ($\beta = 0.43$, S.E. = 0.06, $p = 5.96 \times 10^{-13}$), and negative associations with friendship satisfaction ($\beta = -0.23$, S.E. = 0.06, $p = 4.02 \times 10^{-5}$), neuroticism ($\beta = -0.23$, S.E. = 0.04, $p = 1.67 \times 10^{-9}$), depression ($\beta = -0.04$, S.E. = 0.01, $p = 0.0003$), and cardiovascular problems ($\beta = -0.07$, S.E. = 0.01, $p = 9.32 \times 10^{-6}$). All positive control outcomes, and the negative control outcomes height and height at age 10 were significantly associated with school-leaving age. For comparison, we also associated the outcomes with age at which one left full-time education in regular OLS regression. We did so to examine if analyses that do not take into account causality through a genetic instrument would indicate an association. When doing so, all outcomes except happiness, meaning in life, and bipolar disorder were significantly associated.

Mendelian randomization in sibships

The results from the MR analyses within sibships can be found in Table 6. While our instrument was much less powerful than in regular MR, all *F*-statistics except the *F*-statistic for meaning in life were higher than 10 (which is commonly used as a rule of thumb to avoid bias [Lawlor, Harbord, Sterne, Timpson, & Smith, 2008]). None of the associations (for both control and outcome variables) were significant after correcting for multiple testing.

Interpretation of results

An overview of all results from the four different methods can be found in Fig. 2. As mentioned in the methods section, we define an unambiguous result as one which is consistent across all methods. Additionally, due to the lower power associated with our within-sibship MR analyses, we are satisfied if the magnitude and direction of the Mendelian randomization within siblings is consistent with the other methods.

The ROSLA estimates displayed in Fig. 2 reflect the associations where year of birth was included as a covariate (as preregistered). With respect to our main outcomes, we found non-significant associations across all four methods for: happiness, family satisfaction, work satisfaction, meaning in life, anxiety, and bipolar disorder. Educational duration was positively associated with financial satisfaction, and negatively associated with

https://doi.org/10.1017/S003329172300329X Published online by Cambridge University Press

Table 5. Results Mendelian randomization analyses

	Education (fitted)			*F*-test (1st stage)		Wu–Hausman		Regular OLS education	
Main outcomes	β (S.E.)	*p*	*N*	*F*	*p*	*wh*	*p*	β (S.E.)	*p*
Happiness	−0.04 (0.05)	0.437	185 541	554.6	$<2.2 \times 10^{-16}$	0.472	0.492	−0.004 (0.002)	0.135
Health satisfaction	0.37 (0.05)	**1.31×10^{-14}**	185 706	561.2	$<2.2 \times 10^{-16}$	38.5	5.58×10^{-10}	0.09 (0.003)	**$<2.2 \times 10^{-16}$**
Family satisfaction	−0.05 (0.06)	0.41	134 762	371.3	$<2.2 \times 10^{-16}$	0.185	0.667	−0.02 (0.003)	**5.35×10^{-14}**
Financial satisfaction	0.43 (0.06)	**5.96×10^{-13}**	135 486	373.7	$<2.2 \times 10^{-16}$	28.1	1.15×10^{-7}	0.13 (0.003)	**$<2.2 \times 10^{-16}$**
Friendship satisfaction	−0.23 (0.06)	**4.02×10^{-5}**	134 509	380.2	$<2.2 \times 10^{-16}$	11.2	8×20^{-4}	−0.05 (0.003)	**$<2.2 \times 10^{-16}$**
Work satisfaction	0.07 (0.07)	0.291	89 430	240.1	$<2.2 \times 10^{-16}$	0.58	0.447	0.02 (0.004)	**7.60×10^{-9}**
Meaning in life	−0.12 (0.07)	0.079	103 494	281.5	$<2.2 \times 10^{-16}$	3.49	0.062	0.007 (0.004)	0.036
Neuroticism	−0.23 (0.04)	**1.67×10^{-9}**	262 884	801.3	$<2.2 \times 10^{-16}$	12.1	5.06×10^{-4}	−0.10 (0.002)	**$<2.2 \times 10^{-16}$**
Depression	−0.04 (0.01)	**0.0003**	321 506	957.1	$<2.2 \times 10^{-16}$	5.55	0.018	−0.01 (0.0006)	**$<2.2 \times 10^{-16}$**
Anxiety	0.007 (0.007)	0.333	321 506	957.1	$<2.2 \times 10^{-16}$	3.57	0.059	−0.006 (0.0004)	**$<2.2 \times 10^{-16}$**
Bipolar or manic disorder	0.008 (0.002)	0.734	321 506	957.1	$<2.2 \times 10^{-16}$	0.067	0.795	−0.0002 (0.0001)	0.148
Cardiovascular problems	−0.07 (0.01)	**9.32×10^{-6}**	321 506	957.1	$<2.2 \times 10^{-16}$	5.35	0.021	−0.003 (0.0008)	**$<2.2 \times 10^{-16}$**
Control outcomes									
Income over 18k	0.25 (0.02)	**$<2.2 \times 10^{-16}$**	282 600	797.3	$<2.2 \times 10^{-16}$	91.7	$<2.2 \times 10^{-16}$	0.10 (0.0008)	**$<2.2 \times 10^{-16}$**
Income over 31k	0.29 (0.02)	**$<2.2 \times 10^{-16}$**	282 600	797.3	$<2.2 \times 10^{-16}$	89.4	$<2.2 \times 10^{-16}$	0.13 (0.0009)	**$<2.2 \times 10^{-16}$**
Income over 52k	0.24 (0.02)	**$<2.2 \times 10^{-16}$**	282 600	797.3	$<2.2 \times 10^{-16}$	85.0	$<2.2 \times 10^{-16}$	0.11 (0.0008)	**$<2.2 \times 10^{-16}$**
Income over 100k	0.09 (0.01)	**$<2.2 \times 10^{-16}$**	282 600	797.3	$<2.2 \times 10^{-16}$	58.3	2.21×10^{-14}	0.003 (0.0004)	**$<2.2 \times 10^{-16}$**
Birthweight	0.15 (0.05)	**0.001**	159 371	541.7	$<2.2 \times 10^{-16}$	6.70	0.010	0.03 (0.003)	**$<2.2 \times 10^{-16}$**
Height	0.17 (0.02)	**1.48×10^{-12}**	320 795	954.1	$<2.2 \times 10^{-16}$	12.5	4.00×10^{-4}	0.08 (0.001)	**$<2.2 \times 10^{-16}$**
Comparative body size at age 10	0.01 (0.02)	0.541	312 177	966.2	$<2.2 \times 10^{-16}$	0.397	0.528	−0.001 (0.001)	0.483
Comparative height size at age 10	0.10 (0.02)	**2.03×10^{-5}**	313 496	930.9	$<2.2 \times 10^{-16}$	8.69	0.003	0.03 (0.001)	**$<2.2 \times 10^{-16}$**

Note. Sex, family size, season of birth, year of birth, assessment center, batch, and the first 10 genomic PCs were included as covariates for the MR analyses. The OLS regression is the prediction of the outcomes with education including the same covariates, with the exception of batch and the genomic PCs. All **continuous** outcomes and age at which one left full-time education were standardized.
p-values indicated in bold are lower than the conservative *p*-value threshold of 0.0008.

Table 6. Results Mendelian randomization analyses within sibships

	Education deviation (fitted)			*F*-test (1st stage)		Wu–Hausman	
Main outcomes	β (S.E.)	*p*	*N*	*F*	*p*	*wh*	*p*
Happiness	−0.06 (0.21)	0.778	17 067	22.5	2.15×10^{-6}	0.149	0.699
Health satisfaction	0.001 (0.21)	0.998	17 092	22.7	1.88×10^{-6}	0.002	0.960
Family satisfaction	−0.29 (0.22)	0.179	12 133	21.4	3.70×10^{-6}	4.25	0.039
Financial satisfaction	0.20 (0.22)	0.373	12 128	21.6	3.31×10^{-6}	1.54	0.214
Friendship satisfaction	−0.16 (0.22)	0.463	12 038	21.1	4.38×10^{-6}	1.13	0.287
Work satisfaction	−0.02 (0.25)	0.936	7975	10.1	9.15×10^{-4}	0.005	0.944
Meaning in life	−0.13 (0.49)	0.790	9881	5.08	0.024	0.218	0.640
Neuroticism	0.29 (0.21)	0.162	24 676	27.1	1.24×10^{-7}	0.575	0.016
Depression	−0.002 (0.06)	0.978	30 188	28.2	1.08×10^{-7}	0.0004	0.983
Anxiety	−0.02 (0.04)	0.684	30 188	28.2	1.08×10^{-7}	0.509	0.475
Bipolar or manic disorder	0.04 (0.03)	0.166	30 188	28.2	1.08×10^{-7}	17.2	0.00003
Cardiovascular problems	−0.17 (0.09)	0.074	30 188	28.2	1.08×10^{-7}	9.01	0.003
Control outcomes							
Income over 18k	0.03 (0.09)	0.717	26 674	22.8	1.79×10^{-6}	0.029	0.864
Income over 31k	0.13 (0.10)	0.168	26 674	22.8	1.79×10^{-6}	2.62	0.105
Income over 52k	0.15 (0.06)	0.089	26 674	22.8	1.79×10^{-6}	5.40	0.020
Income over 100k	−0.01 (0.04)	0.799	26 674	22.8	1.79×10^{-6}	0.31	0.584
Birthweight	−0.15 (0.23)	0.504	15 048	16.4	5.05×10^{-5}	0.984	0.321
Height	−0.05 (0.10)	0.624	30 136	29.2	6.53×10^{-8}	0.471	0.492
Comparative body size at age 10	0.17 (0.14)	0.226	29 337	26.8	2.27×10^{-7}	3.30	0.069
Comparative height size at age 10	0.11 (0.12)	0.390	29 375	24.8	6.39×10^{-7}	1.27	0.260

p-values indicated in bold are lower than the conservative *p*-value threshold of 0.0008.

neuroticism in the sibling-control and MR analyses, but these associations were non-significant in both the ROSLA and within-sibling MR. The within-sibling MR estimate for financial satisfaction was in the same direction, and of comparable magnitude as the conventional MR results, whereas the result for neuroticism was in the opposite direction. Lastly, educational duration was significantly positively associated with health satisfaction and significantly negatively associated with friendship satisfaction, depression, and cardiovascular outcomes in the conventional MR analyses only. Thus, overall, the different analyses do not seem to converge on a consistent conclusion in terms of whether there is a causal effect.

We included different income classes as positive control outcomes, as we expected educational duration to causally influence income. Only in the sibling control and MR analyses were the different income variables significantly associated with educational duration. The non-significant within-family MR estimates for income were in the same direction but of slightly smaller magnitude as the conventional MR.

When not including year of birth as a covariate in the ROSLA analyses, the first three income classes were significantly associated with educational duration, suggesting a potential overcorrection in our ROSLA analyses (online Supplementary Table S2). With respect to our negative controls, height was significantly associated with education in both the sibling-control and MR analyses. Additionally, comparative body height at age 10 was significantly associated with education in the MR analyses. Associations with these negative control phenotypes suggest the possible presence of residual bias.

Discussion

Our study was designed to disentangle causal effects from confounding in the association between educational duration and different well-being, and mental and physical health indicators. To this end, we applied four established techniques for causal inference to a homogeneous sample, the UKB. We find consistent non-significant associations for happiness, family satisfaction, work satisfaction, meaning in life, depression, anxiety, and bipolar disorder. However, we do not find robust significant associations across all four methods for health satisfaction, friendship satisfaction, financial satisfaction, neuroticism, and cardiovascular outcomes. The absence of significant consistent results suggests that associations between educational duration and well-being, mental and physical health are largely confounded or biased by reverse causation. Alternatively, a small causal effect may exist but power in one or some of our techniques may have been insufficient to detect it.

Overall, in our first set of analyses (based on the ROSLA), we do not find an effect of educational duration on any of the outcomes, including our positive controls. This contradicts an earlier

https://doi.org/10.1017/S003329172300329X Published online by Cambridge University Press

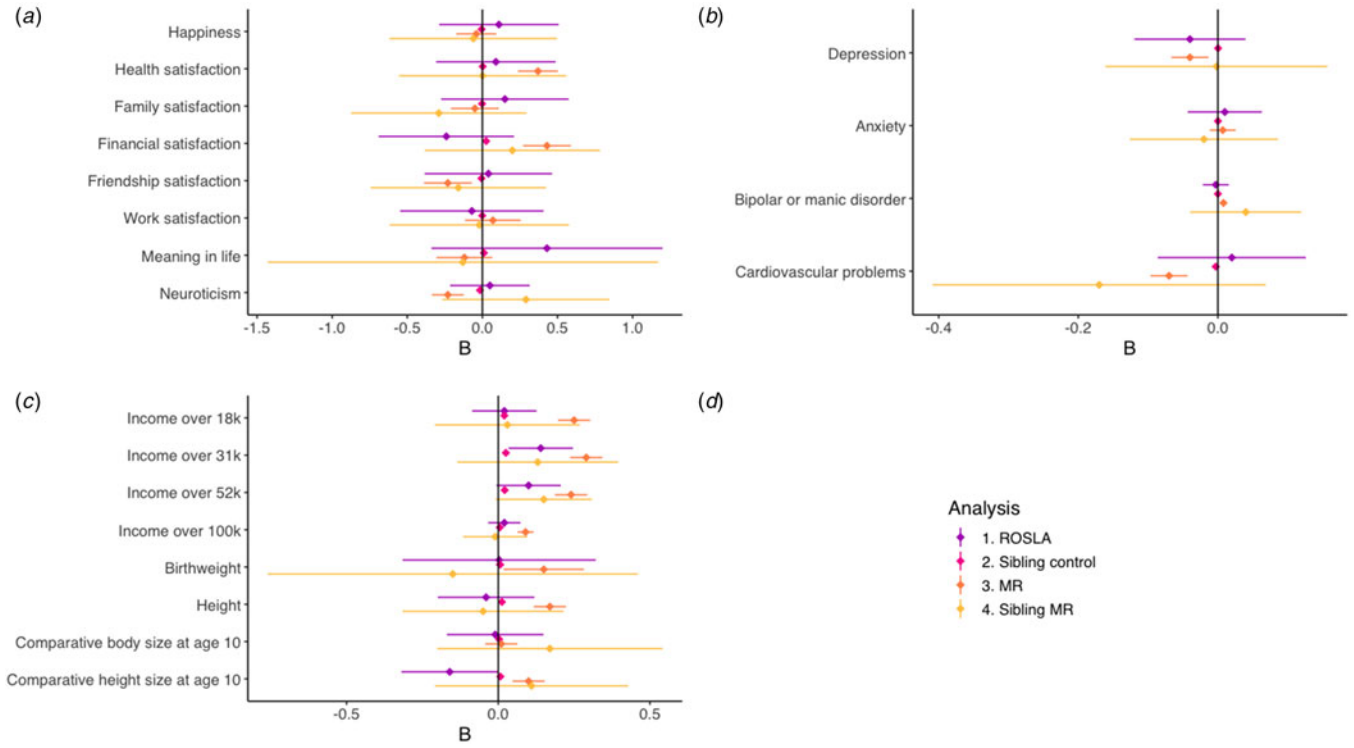


Figure 2. Overview of the results from the different analyses for (a) continuous outcome measures, (b) binary outcome measures, and (c) control measures.

study similarly examining the causal effect of education in the UK in light of the ROSLA reform, where a causal effect was found for different cardiovascular outcomes, income, and height. The main difference between the current study and the Davies et al. study is the method of correcting for year of birth. Whereas Davies et al. employ a difference-in-difference approach where the data are stratified by year of birth, we directly include year of birth as a covariate. To examine the effect of year of birth on the associations, we ran supplementary analyses where we compare the associations with and without year of birth as a covariate. When including year of birth, educational duration no longer significantly influenced our positive control outcomes (the four income classes), suggesting that by including year of birth in this set of analyses we might be overcorrecting. This degree of sensitivity of the model to inclusion of covariates, and to what way covariates are taken into account, does complicate the interpretation of our findings.

Although the ROSLA analyses did not result in significant associations, we found an effect of educational duration on health satisfaction, friendship satisfaction, depression, and cardiovascular outcomes, but only in the conventional (non-family) MR analyses. The finding that these associations were only significant in the MR analyses might indicate that one of the assumptions was violated, such as the no pleiotropy assumption. Moreover, we found significant associations with financial satisfaction and neuroticism in both the conventional MR and within-sibship analyses. For the latter, we found that people who stayed in school longer had higher financial satisfaction and lower neuroticism. While these associations were not significant in the within-sibling MR analyses, the direction of effect was consistent. We can therefore interpret these associations as at least suggestive evidence in three out of four analyses. Besides our potential overcorrection issue, it could be argued that discrepancies with the ROSLA analyses are caused by the caveat that the ROSLA results only apply to those who would have left school at age 15 in the absence of the reform. In this sense, the ROSLA results are less generalizable to the population than the other methods as the reform did not affect those who would have stayed in school until age 16 or later irrespective of the reform. Additionally, one of the possibilities that comes to mind when interpreting the results for financial satisfaction is the potential mediation of income. We therefore re-ran the sibling control analyses including the mean income and the sibling deviation in income from the sibship income average as covariates (Saunders, McGue, & Malone, 2019) (see online Supplementary Table S3). We found that the standardized estimates for both financial satisfaction and neuroticism decreased substantially (with similar standard errors), resulting in non-significant associations controlling for income, suggesting that the association between educational duration and these two outcomes may be mediated through income. Since financial satisfaction is partly the result of one's income, this finding is not surprising. With respect to neuroticism, a previous MR study found evidence for bidirectional causality between education and neuroticism, but did not consider potential mediation of income (Nagel et al., 2018).

The most consistent finding to emerge from the data is the lack of evidence for a causal effect of educational duration on happiness, family satisfaction, work satisfaction, meaning in life, anxiety, and bipolar disorder. The association between educational duration and these outcomes was non-significant, irrespective of which method was applied, while OLS suggested there was an association. In these OLS analyses, we observe that almost all outcomes were significantly predicted by school-leaving age. It is therefore likely that the OLS associations are subject to confounding and/or reverse causation, and are unlikely to reflect direct, causal effects. One note in the context of our findings is that

https://doi.org/10.1017/S003329172300329X Published online by Cambridge University Press

we examine variation in educational duration from a minimal school-leaving age onward, and do not examine the effect of attending education in general. It is therefore important to note that schooling in general has important pecuniary and non-pecuniary consequences (Grossman, 2006), but that our study suggests a lack of evidence for causal effects of variation in educational duration on variation in (mental) health outcomes beyond a minimum school-leaving age. Additionally, important to note is that the current project focused on potential causal effects of educational duration on (mental) health, and not reverse effect or bi-directional causality. It might be the case that, for some phenotypes, there is a causal effect from health on educational duration. In addition, we focused on linear associations, disregarding potential non-linear effects. While these two directions were beyond the scope of the current project, they would be interesting directions for future research.

When applying the methods used here in isolation, it is often difficult or impossible to evaluate all the respective limitations and assumptions. A strength of this study is that we try to minimize our reliance on any one set of assumptions by applying various existing approaches for causal inference that rely on different assumptions to account for possible confounding and bias, and triangulate results. In doing so, we found that the different causal inference approaches led to heterogeneous results. Since we investigate the same measures in (largely) the same population, differences in results across methods are most likely attributable to the methods themselves. Importantly, if we decided to focus on only one of these methods for the current paper, we would have drawn very different conclusions than we do now. With respect to health satisfaction, friendship satisfaction, and cardiovascular outcomes, these were only significantly predicted by educational duration in the conventional MR analyses, but not in any of the other analyses. Evaluating this discrepancy considering the characteristics of the different methods, it is possible that these associations are caused by a familial or population effect that is uncontrolled for in conventional MR but is controlled for in the other analyses. Additionally, the MR sample was the largest sample we examined, and it might be the case that an increase in sample size for the ROSLA and within-sibship analyses would allow us to detect smaller effects that remain undetected using the current sample size. This is also reflected in the confidence intervals for the results from the different methods (Fig. 2), where we see relatively precise estimates for the sibling control and MR analyses, and relatively imprecise estimates for the within-sibling MR and especially the ROSLA analyses. Therefore, it might be the case that the lack of significant results in the ROSLA and within-sibling MR is due to a lack of power. Regardless of power, the negative control traits suggest a reliance on MR alone risks false-positive results for obvious reasons: the significant causal effects of educational duration on birthweight and height at age 10 cannot be true effects.

While we tried to account for the limitations of the separate methods by means of triangulation, our results are still sensitive to our sample and measurement characteristics. We used a relatively homogeneous sample that allowed for a straightforward comparison between methods, but this also limits the generalizability of our findings. More specifically, the UKB sample is known to suffer from a 'healthy volunteer' bias, where participants are more healthy than the general population (Fry et al., 2017; Munafò, Tilling, Taylor, Evans, & Davey Smith, 2018). Additionally, participants are more likely to be older, female, and live in more socioeconomically advantaged areas than non-participants (Fry et al., 2017). To (partly) correct for the confounding stemming from volunteer bias, we used weights as calculated in van Alten et al. (2022). While these weights help us correct for volunteer bias, it is important to stress that the examined sample is still a WEIRD (Western, Educated, Industrialized, Rich, Democratic) sample. It is therefore unknown if these results generalize to non-WEIRD societies. We also tested if the conclusions of the sibling control analyses would change if we did not restrict to European Ancestry individuals born in England and Wales. We did not find different results in this larger set of siblings (see online Supplementary Table 4). Second, we used relatively broad, imprecise phenotype and disease definitions. For example, we included all depression diagnoses present in UKB under the umbrella 'depression', and all cardiovascular-related diagnoses under the umbrella 'cardiovascular outcomes'. For our continuous phenotypes (except neuroticism), we used single items to measure the phenotypes. It is possible that more precise phenotype and disease definitions could reduce measurement error and influence power. Additionally, it has been argued that quantitative education measures such as years of education is not an optimal measure of education, especially in the context of non-pecuniary returns of education (Oreopoulos & Salvanes, 2009). More qualitative measures of education, such as teaching methods or curricula differences, might be better suited in this context, but these data are difficult to acquire and analyze on a large scale. Our quantitative measure of education was moreover collected through self-report questionnaires, and those responses might have suffered from recall bias (i.e. participants might not remember school-leaving age accurately). We did use single items for our well-being phenotypes, but we did not treat well-being as a unidimensional construct. Alternative to looking at a general well-being item or sum-score, we assessed if educational duration influenced specific well-being aspects, such as work satisfaction and meaning in life. However, these well-being measures also relied on self-report, which might have introduced measurement error. When examining the test–retest reliability for education and happiness, which were measured on multiple occasions, we find moderate test–retest reliabilities between 0.57 and 0.68, which is in line with previous findings on the stability of well-being (Anusic & Schimmack, 2016). Lastly, the included analyses do not necessarily estimate the same type of causal effect. While the ROSLA and conventional MR identify LATE effects, the sibling analyses identify Conditional Average Treatment Effects (CATE). For ROSLA and MR, this means that we estimate an effect in those whose treatment (educational duration) would differ if the value of the instrumental variable differed. However, note that under different assumptions, an MR effect can be identified as an Average Treatment Effect (ATE). An MR effect can be identified as ATE when we assume the exposure (i.e. education) has the same causal effect on the outcomes for the whole population, or when the instrument (i.e. genetic variants associated with education) has a consistent effect on the exposure, and that this effect is the same for the whole population. In addition, ATE can be identified under the no simultaneous heterogeneity assumption (NoSH), which assumes that the heterogeneity in the instrument–exposure effect and the exposure–outcome effect should be uncorrelated (Hartwig, Wang, Davey Smith, & Davies, 2023). For CATE, the causal interpretation we obtain only applies to our sub-population of sibling pairs, excluding singletons. In addition, if the exposure of one sibling affects the exposure of the other, the estimator describes what happens if the treatment affects both siblings (Petersen & Lange, 2020).

https://doi.org/10.1017/S003329172300329X Published online by Cambridge University Press

We used a natural experiment, sibling-control analysis, Mendelian randomization, and within-sibship Mendelian randomization to a large UK sample to disentangle potential causal effects of education duration on several mental and physical health outcomes. A comparison of results across these four methods illustrates that (1) associations between education and these several outcomes are largely confounded, and (2) triangulation of evidence across different methods is necessary to examine the results in light of their respective limitations. Notwithstanding the relatively limited generalizability of our findings across different cultures, time frames, and educational systems, this work provides valuable insight into the complexities of establishing the causal effects of EA on important life outcomes.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S003329172300329X.

Acknowledgements. This research has been conducted using the UK Biobank Resource under Application Number 40310.

Author contributions. All authors contributed to the study conception and design. Data analyses were performed by Margot van de Weijer, Perline Demange, and Michel Nivard. The first draft of the manuscript was written by Margot van de Weijer and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding statement. Margot van de Weijer, Dirk Pelt, and Meike Bartels are supported by the European Research Council Consolidator Grant (ERC-2017-COG 771057 WELL-BEING PI Bartels). Margot van de Weijer is funded by the European Union (ERC, UNRAVEL-CAUSALITY, project nr. 101076686). Views and opinions expressed are however those of the author (s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. Perline Demange is supported by the grant 531003014 from The Netherlands Organisation for Health Research and Development (ZonMW). Michel Nivard is supported by R01MH120219, ZonMW grants 849200011919 and 531003014 from The Netherlands Organisation for Health Research and Development, a VENI grant awarded by NWO (VI.Veni.191G.030), and is a Jacobs Foundation Research Fellow.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Ethical standards. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

References

Anusic, I., & Schimmack, U. (2016). Stability and change of personality traits, self-esteem, and well-being: Introducing the meta-analytic stability and change model of retest correlations. *Journal of Personality and Social Psychology*, *110*(5), 766–781. https://doi.org/10.1037/pspp0000066.

Banks, J., & Mazzonna, F. (2012). The effect of education on old age cognitive abilities: Evidence from a regression discontinuity design. *The Economic Journal*, *122*(560), 418–448. https://doi.org/10.1111/J.1468-0297.2012.02499.X.

Batty, G. D., Gale, C. R., Kivimäki, M., Deary, I. J., & Bell, S. (2020). Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: Prospective cohort study and individual participant meta-analysis. *BMJ*, *368*, m131. https://doi.org/10.1136/bmj.m131

Bergé, L. (2018). Efficient estimation of maximum likelihood models with multiple fixed-effects: The R package FENmlm. CREA Discussion Papers, 13.

Brumpton, B., Sanderson, E., Heilbron, K., Hartwig, F. P., Harrison, S., Vie, G. Å., … Davies, N. M. (2020). Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses. *Nature Communications*, *11*(1), 3519. https://doi.org/10.1038/s41467-020-17117-4

Brunello, G., Fort, M., & Weber, G. (2009). Changes in compulsory schooling, education and the distribution of wages in Europe. *The Economic Journal*, *119*(536), 516–539. https://doi.org/10.1111/j.1468-0297.2008.02244.x

Bücker, S., Nuraydin, S., Simonsmeier, B. A., Schneider, M., & Luhmann, M. (2018). Subjective well-being and academic achievement: A meta-analysis. *Journal of Research in Personality*, *74*, 83–94. https://doi.org/10.1016/J.JRP.2018.02.007

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., … Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Chevalier, A., & Feinstein, L. (2006). Sheepskin or prozac: The causal effect of education on mental health. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.923530

Choi, A. I., Weekley, C. C., Chen, S. C., Li, S., Kurella Tamura, M., Norris, K. C., & Shlipak, M. G. (2011). Association of educational attainment with chronic disease and mortality: The kidney early evaluation program (KEEP). *American Journal of Kidney Diseases*, *58*(2), 228–234. https://doi.org/10.1053/J.AJKD.2011.02.388

Choi, S. W., Mak, T. S.-H., & O'Reilly, P. F. (2020). Tutorial: A guide to performing polygenic risk score analyses. *Nature Protocols*, *15*(9), 2759–2772. https://doi.org/10.1038/s41596-020-0353-1

Clark, D., & Royer, H. (2013). The effect of education on adult mortality and health: Evidence from Britain. *American Economic Review*, *103*(6), 2087–2120. https://doi.org/10.1257/aer.103.6.2087

Davies, N. M., Dickson, M., Davey Smith, G., Windmeijer, F., & van den Berg, G. J. (2019a, May 21). The causal effects of education on adult health, mortality and income: Evidence from Mendelian randomization and the raising of the school leaving age [IZA Discussion Paper No. 12192]. Rochester, NY. https://doi.org/10.2139/ssrn.3390179

Davies, N. M., Dickson, M., Davey Smith, G., Windmeijer, F., & van den Berg, G. J. (2021). The causal effects of education on adult health, mortality and income: Evidence from Mendelian randomization and the raising of the school leaving age. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.3390179

Davies, N. M., Dickson, M., Smith, G. D., Van Den Berg, G. J., & Windmeijer, F. (2018). The causal effects of education on health outcomes in the UK Biobank. *Nature Human Behaviour 2017*(2), 2, 2(2), 117–125. https://doi.org/10.1038/S41562-017-0279-Y

Davies, N. M., Hill, W. D., Anderson, E. L., Sanderson, E., Deary, I. J., & Davey Smith, G. (2019b). Multivariable two-sample Mendelian randomization estimates of the effects of intelligence and education on health. *eLife*, *8*, e43990. https://doi.org/10.7554/eLife.43990

Fewell, Z., Davey Smith, G., & Sterne, J. A. C. (2007). The impact of residual and unmeasured confounding in epidemiologic studies: A simulation study. *American Journal of Epidemiology*, *166*(6), 646–655. https://doi.org/10.1093/aje/kwm165

Frisell, T. (2021). Invited commentary: Sibling-comparison designs, are they worth the effort? *American Journal of Epidemiology*, *190*(5), 738–741. https://doi.org/10.1093/aje/kwaa183

Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., … Ellen, N. E. (2017). Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. *American Journal of Epidemiology*, *186*(9), 1026–1034.

Furnée, C. A., Groot, W., & Van Den Brink, H. M. (2008). The health effects of education: A meta-analysis. *European Journal of Public Health*, *18*(4), 417–421. https://doi.org/10.1093/EURPUB/CKN028

Gill, D., Efstathiadou, A., Cawood, K., Tzoulaki, I., & Dehghan, A. (2019). Education protects against coronary heart disease and stroke independently of cognitive function: Evidence from Mendelian randomization. *International Journal of Epidemiology*, *48*(5), 1468–1477. https://doi.org/10.1093/ije/dyz200

Glymour, M. M., & Manly, J. J. (2018). Compulsory schooling laws as quasi-experiments for the health effects of education: Revisiting theory to understand mixed results. *Social Science & Medicine* (1982), *214*, 67–69. https://doi.org/10.1016/j.socscimed.2018.08.008

Graeber, D. (2017). Does more education protect against mental health problems? (Research Report No. 113). DIW Roundup: Politik im Fokus. Retrieved from DIW Roundup: Politik im Fokus website Retrieved from https://www.econstor.eu/handle/10419/169442

Grenet, J. (2013). Is extending compulsory schooling alone enough to raise earnings? Evidence from French and British compulsory schooling laws*. *The Scandinavian Journal of Economics*, *115*(1), 176–210. https://doi.org/10.1111/J.1467-9442.2012.01739.X

Grossman, M. (2006). Chapter 10 education and nonmarket outcomes. In E. Hanushek & F. Welch (Eds.), *Handbook of the economics of education* (Vol. *1*, pp. 577–633). Amsterdam: North-Holland, the Netherlands: Elsevier. https://doi.org/10.1016/S1574-0692(06)01010-5

Hamad, R., Elser, H., Tran, D. C., Rehkopf, D. H., & Goodman, S. N. (2018). How and why studies disagree about the effects of education on health: A systematic review and meta-analysis of studies of compulsory schooling laws. *Social Science & Medicine*, *212*, 168–178. https://doi.org/10.1016/J.SOCSCIMED.2018.07.016

Hamad, R., Nguyen, T. T., Bhattacharya, J., Glymour, M. M., & Rehkopf, D. H. (2019). Educational attainment and cardiovascular disease in the United States: A quasi-experimental instrumental variables analysis. *PLoS Medicine*, *16*(6), e1002834. https://doi.org/10.1371/journal.pmed.1002834

Hanscombe, K. B., Coleman, J. R. I., Traylor, M., & Lewis, C. M. (2019). ukbtools: An R package to manage and query UK Biobank data. *PLoS ONE*, *14* (5), e0214311. https://doi.org/10.1371/JOURNAL.PONE.0214311

Hartwig, F. P., Wang, L., Davey Smith, G., & Davies, N. M. (2023). Average causal effect estimation via instrumental variables: The no simultaneous heterogeneity assumption. *Epidemiology*, *34*(3), 325–332. https://doi.org/10.1097/EDE.0000000000001596

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., … Haycock, P. C. (2018). The MR-Base platform supports systematic causal inference across the human phenome. *eLife*, *7*, e34408. https://doi.org/10.7554/eLife.34408

Hothorn, T., Zeileis, A., Farebrother (pan.f), R. W., Cummins (pan.f), C., Millo, G., & Mitchell, D. (2022). lmtest: Testing linear regression models. Retrieved from https://cran.r-project.org/web/packages/lmtest/index.html

Howe, L. D., Kanayalal, R., Harrison, S., Beaumont, R. N., Davies, A. R., Frayling, T. M., … Tyrrell, J. (2020). Effects of body mass index on relationship status, social contact and socio-economic position: Mendelian randomization and within-sibling study in UK Biobank. *International Journal of Epidemiology*, *49*(4), 1173–1184. https://doi.org/10.1093/ije/dyz240

Huang, C. (2015). Academic achievement and subsequent depression: A meta-analysis of longitudinal studies. *Journal of Child and Family Studies*, *24*(2), 434–442. https://doi.org/10.1007/S10826-013-9855-6/TABLES/2

Ichino, A., & Winter-Ebmer, R. (1999). Lower and upper bounds of returns to schooling: An exercise in IV estimation with different instruments. *European Economic Review*, *43*(4), 889–901. https://doi.org/10.1016/S0014-2921(98)00102-0

Kawachi, I., Adler, N. E., & Dow, W. H. (2010). Money, schooling, and health: Mechanisms and causal evidence. *Annals of the New York Academy of Sciences*, *1186*, 56–68. https://doi.org/10.1111/j.1749-6632.2009.05340.x

Khaing, W., Vallibhakara, S. A., Attia, J., McEvoy, M., & Thakkinstian, A. (2017). Effects of education and income on cardiovascular outcomes: A systematic review and meta-analysis. *European Journal of Preventive Cardiology*, *24*(10), 1032–1042. https://doi.org/10.1177/2047487317705916

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., & Smith, G. D. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, *27*(8), 1133–1163. https://doi.org/10.1002/SIM.3034

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., … Cesarini, D. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, *50*(8), 1112–1121. https://doi.org/10.1038/s41588-018-0147-3

Lleras-Muney, A. (2002). Were compulsory attendance and child labor laws effective? An analysis from 1915 to 1939. *The Journal of Law and Economics*, *45*(2), 401–435. https://doi.org/10.1086/340393

Lorant, V., Deliège, D., Eaton, W., Robert, A., Philippot, P., & Ansseau, M. (2003). Socioeconomic inequalities in depression: A meta-analysis. *American Journal of Epidemiology*, *157*(2), 98–112. https://doi.org/10.1093/AJE/KWF182

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873. https://doi.org/10.1093/BIOINFORMATICS/BTQ559

Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M., & Davey Smith, G. (2018). Collider scope: When selection bias can substantially influence observed associations. *International Journal of Epidemiology*, *47*(1), 226–235. https://doi.org/10.1093/ije/dyx206

Nagel, M., Jansen, P. R., Stringer, S., Watanabe, K., de Leeuw, C. A., Bryois, J., … Posthuma, D. (2018). Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nature Genetics*, *50*(7), 920–927. https://doi.org/10.1038/s41588-018-0151-7

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., … Benjamin, D. J. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 2016 *533*:7604, 533 (7604), 539–542. https://doi.org/10.1038/NATURE17671

Oreopoulos, P., & Salvanes, K. G. (2009). How large are returns to schooling? Hint: Money isn't everything. In NBER Working Papers (No. 15339). National Bureau of Economic Research, Inc. Retrieved from National Bureau of Economic Research, Inc website: https://ideas.repec.org/p/nbr/nberwo/15339.html

Petersen, A. H., & Lange, T. (2020). What is the causal interpretation of sibling comparison designs? *Epidemiology*, *31*(1), 75–81. https://doi.org/10.1097/EDE.0000000000001108

Plotnikov, D., Williams, C., Atan, D., Davies, N. M., Mojarrad, N. G., & Guggenheim, J. A. (2020). Effect of education on myopia: Evidence from the United Kingdom ROSLA 1972 reform. *Investigative Ophthalmology & Visual Science*, *61*(11), 7–7. https://doi.org/10.1167/IOVS.61.11.7

Putrik, P., Ramiro, S., Keszei, A. P., Hmamouchi, I., Dougados, M., Uhlig, T., … Boonen, A. (2016). Lower education and living in countries with lower wealth are associated with higher disease activity in rheumatoid arthritis: Results from the multinational COMORA study. *Annals of the Rheumatic Diseases*, *75*(3), 540–546. https://doi.org/10.1136/ANNRHEUMDIS-2014-206737

Saunders, G. R. B., McGue, M., & Malone, S. M. (2019). Sibling comparison designs: Addressing confounding bias with inclusion of measured confounders. *Twin Research and Human Genetics*, *22*(5), 290–296. https://doi.org/10.1017/thg.2019.67

Schimmack, U. (2008). The structure of subjective well-being. In M. Eid & R. J. Larsen (Eds.), *The science of subjective well-being* (pp. 97–123). New York, NY: Guilford Press.

Sobel, M. E. (2000). Causal inference in the social sciences. *Journal of the American Statistical Association*, *95*(450), 647–651. https://doi.org/10.2307/2669410

Telfair, J., & Shelton, T. L. (2012). Educational attainment as a social determinant of health. *North Carolina Medical Journal*, *73*(5), 358–365. https://doi.org/10.18043/NCM.73.5.358

van Alten, S., Domingue, B. W., Galama, T., & Marees, A. T. (2022, May 16). Reweighting the UK Biobank to reflect its underlying sampling population substantially reduces pervasive selection bias due to volunteering (p. 2022.05.16.22275048). p. 2022.05.16.22275048. medRxiv. https://doi.org/10.1101/2022.05.16.22275048

van der Heide, I., Wang, J., Droomers, M., Spreeuwenberg, P., Rademakers, J., & Uiters, E. (2013). The relationship between health, education, and health literacy: Results from the Dutch adult literacy and life skills survey. *Journal of Health Communication*, *18*(sup1), 172–184. https://doi.org/10.1080/10810730.2013.825668

https://doi.org/10.1017/S003329172300329X Published online by Cambridge University Press