Early and Later-life Stimulation: How Retirement Shapes the Effect of Education on Old-age Cognitive Abilities*

Hendrik Schmitz
Paderborn University,
RWI Essen, and
Leibniz Science Campus Ruhr

Matthias Westphal
TU Dortmund,
RWI Essen, and
Leibniz Science Campus Ruhr

February 2023

Abstract

We study the interaction of education in adolescence and labor force participation around retirement age and its effect on cognitive abilities for individuals in Europe aged 50–70. In addition to a direct long-run effect, indirect ones may arise, specifically through labor force participation. We directly test this using a novel and purpose-built estimator for causal mediation analysis that accommodates endogeneity and heterogeneous treatment effects. Overall, we find that education raises cognitive abilities by about 8 percent. Only among the more educated, 36 percent of this effect can be attributed to labor force participation, emphasizing important complementarities between both factors.

Keywords: Cognitive abilities, causal mediation analysis, marginal treatment effects, education

JEL Classification: *C31, J14, J24*

Hendrik Schmitz: Paderborn University, Warburger Strasse 100, 33098 Paderborn, Germany, Tel.: +49 5251 603213, E-mail: hendrik.schmitz@uni-paderborn.de.

Matthias Westphal: Technische Universität Dortmund, Fakultät Wirtschaftswissenschaften, 44221 Dortmund, Germany, Tel.: +49 231 755-5403, E-mail: matthias.westphal@tu-dortmund.de.

*We thank Jochem de Bresser, Martin Huber, Daniel Kuehnle and the participants at the IAAE annual conference, the EEA annual meeting, the Essen Health Conference, the health and population economics seminar at the Leibniz University Hannover, the annual meeting of the German Health Economics Association, and the International Health Economics Workshop Tilburg for helpful comments and discussion. Financial support by the Deutsche Forschungsgemeinschaft is gratefully acknowledged. Detailed acknowledgements for the data at the end of the text.

1 Introduction

Evidence from neuroscience and economic research let it appear irrevocable that cognitive abilities decline with age (Kaufman and Horn, 1996; Grady, 2012; Strittmatter et al., 2020). This decline has considerable implications for human interactions, economic choices, and the quality of life per se (see, e.g. Tymula et al., 2013, Christelis et al., 2010, Banks and Oldfield, 2007, Banks et al., 2010, Smith et al., 2010). For instance, dementia—referring to a group of symptoms that originates from the most drastic form of cognitive decline—strongly and increasingly affects many parts of society, ranging from families to the health- and long-term-care systems (Chandra et al., 2022). Additionally, neuroscientific associations between individual life styles and cognitive aging suggest that individual behaviors may prevent or at least defer these negative implications (Lindenberger, 2014). Knowledge about the causal determinants of these associations would be key for sustainable aging societies.

In this paper, we study whether and to which extent the decline and its implications are malleable by education—an important decision in life with many downstream implications over the life course. Although the literature on determinants of skill formation seems to be settled on the fact that hardly anything, including education, impacts *adolescent* cognitive abilities after, say, age 10 (see Heckman, 2007, 2008 for overviews and Cornelissen and Dustmann, 2019 for recent evidence), a number of studies does find effects of education on the *old-age* cognitive decline (see Figure 1 below). We argue that because of the downstream implications, such as occupational choices, labor supply and retirement decisions, this seemingly puzzling finding might make sense.

To provide some intuition, there are two broad ways, in which we think education may affect a cognitive decline. First, according to the cognitive reserve hypothesis, individuals with a higher cognitive capacity may be more resilient against an age-related decline (Fratiglioni and Wang, 2007). Thus, if education boosted this capacity, one would expect a slower cognitive decay. Without an effect of education on the level of pre-decline cognitive abilities, however, there would be less scope for this mechanism to work. Second, even if such a direct effect of education on cognitive abilities does not exist, education may still indirectly affect the cognitive decline, because it may change many other dimensions in life. Take, for instance, labor force participation. Education most directly influences this decision through occupational choices and the career starting age. While low-educated individuals may be more likely to do routine or physical work, more educated individuals could be more likely to chose more cognitively-stimulating jobs. Moreover, if total work experience at retirement was unaffected by education, more educated individuals would need to work until older ages, as they enter the labor market later. Both points would imply a different environment of cognitive stimulation of individuals in their late fifties to sixties, which are typical starting ages for a natural and perceptible cognitive decline.

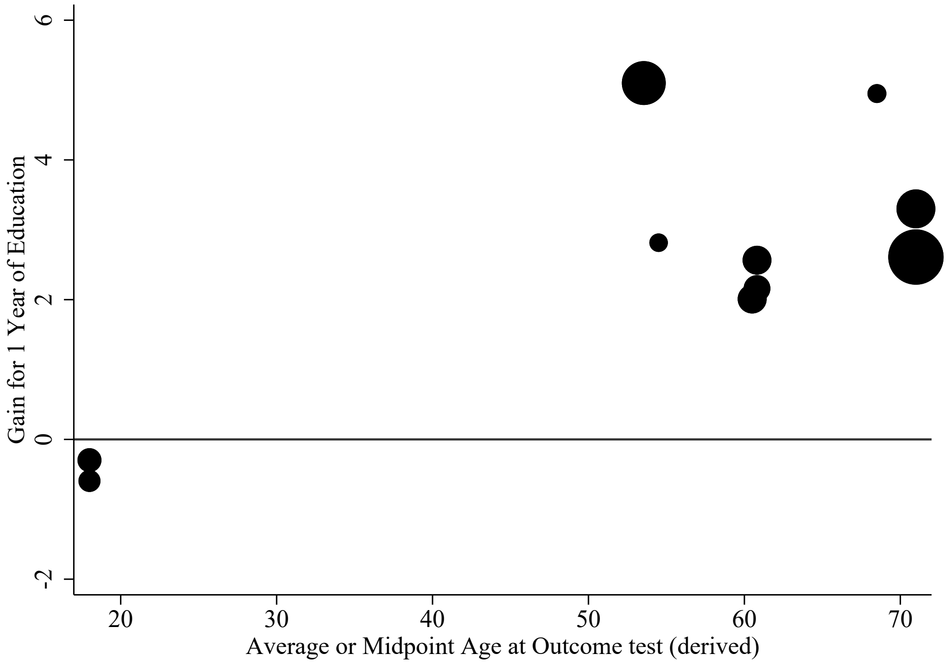
This second view on the effect of education on cognitive decline is boosted by an intriguing observation that summarizes well the literature on cognitive decline. Using the data from the meta analysis by Ritchie and Tucker-Drob (2018) and including only results for fluid intelligence (which is a more sensitive leading indicator of a cognitive decline) from studies in which education is instrumented by some policy change, Figure 1 suggests that the causal effect of education seems to increase with age.[1] Besides these studies on measures on fluid intelligence, Seblova et al. (2021) assess effects on dementia risks directly. They use a sample of very old individuals (with an average age of above 80) and their exposure to a compulsory schooling reform and find that education does not seem to lead to a decreased dementia risk. Thus, there is limited evidence for a direct effect of education (i.e. the cognitive reserve hypothesis does not seem to hold for dementia), but the authors assume that education may alter other risk factors (such as occupational status) in adult life that at least temporarily affect cognitive abilities. We do not claim that this is an exhaustive list of results by age group in the literature, nor that all studies are perfectly comparable with respect to reforms or measures of cognitive abilities. Nevertheless, the effects of education on cognitive abilities in these studies tend to increase in age, in particular in ages where some individuals still work. Because individuals with more education are less likely to retire earlier and retirement itself, as is well-documented, has negative effects on cognition[2], retirement may be the driving force behind Figure 1.

By assessing how retirement empirically alters the effect of education on cognitive decline, we make two main contributions to the literature. First, we develop a novel estimator for causal mediation analyses that is able to quantify the role of labor force participation in the formation of

[1]Studies that use crystallized intelligence scores (capturing acquired knowledge) as outcomes and, again, compulsory schooling reforms as exogenous variation find a precise positive effect around school-leaving age (see, e.g., Brinch and Galloway, 2012), but a more ambiguous effect at later ages (Kamhöfer and Schmitz, 2016; Carlsson et al., 2015; Glymour et al., 2008; Schneeweis et al., 2014). Potentially, this non-amplifying effect along the life can be explained by a decline of the specific knowledge learned in school.

[2]See, for instance, Rohwedder and Willis (2010); Bonsang et al. (2012); Coe et al. (2012); Celidoni et al. (2017); Mazzonna and Peracchi (2012, 2017); Atalay et al. (2019); Schmitz and Westphal (2021).

Figure 1: The association between age and the effect of education on IQ in compulsory schooling studies



Notes: Based on the meta analysis by Ritchie and Tucker-Drob (2018) complemented by the recent study by Hampf (2019). The remaining studies are Banks and Mazzonna (2012); Carlsson et al. (2015); Glymour et al. (2008); Gorman et al. (2017); Schneeweis et al. (2014). The point size is inversely proportional to the standard error of the respective effect. IQ points are measured by a standardized score with a mean of 100 and a standard deviation of 15.

the total effect of education. This estimator is flexible and accommodates endogenous treatment and mediator choices as well as heterogeneous treatment effects. Conventional IV estimation—including treatment, mediator, and their interaction as regressors, as for instance employed in Chen et al. (2019)—does not identify the contribution of the mediator correctly if there is unobserved heterogeneity and effects for individuals who react to the instrument for the treatment (the compliers) differ substantially from individuals who do not (the always takers and never takers). Our general estimator can be applied to all settings in which the aim is to decompose the causal effect of a treatment into a direct effect and an indirect one that runs through the mediator. Except for Frölich and Huber (2017)[3], existing methods either address endogeneity of treatment *or* mediator (as discussed in Keele et al., 2015) but not of both, or need to make comparably strong assumptions, for instance to allow one instrument to jointly solve both endogeneity problems (Dippel et al.,

[3]This approach is also employed by Salm et al. (2021). Chen et al. (2020) adjust this estimator to stochastic frontier models.

2020).[4] Basing our estimator on direct instrumental variables estimation, we complement Frölich and Huber (2017) who use a control function approach. We employ the marginal treatment effect (MTE) framework (Heckman and Vytlacil, 1999), which highlights the role of selection in effect heterogeneity and allows us to present the general idea underlying our approach to mediation analysis intuitively using graphs. The estimation approach also allows to transparently report the mechanics behind our results. Moreover, while Frölich and Huber (2017) present cases for different combinations of binary/continuous mediator and binary/continuous instrument, they do not address the case that applies here: binary treatment, binary mediator, and two binary instruments. By expanding MTE estimation with binary instruments (Brinch et al., 2017) to causal mediation analysis, we make this approach applicable also in this (not so uncommon) case.

The second main contribution of this paper is one of content. We apply our estimator to precisely identify the extent to which labor force participation moderates the effect of education on cognitive decline, thus analyzing an important pathway of the effect of education. While the literature has so far explored labor-force participation as a potential mechanism of the effect of education, this mechanism has not yet been fully quantified. Typically, by also reporting the effect of education on labor-force participation, the studies report indirect evidence—as we will show, this is only a necessary, but by no means sufficient condition of labor-force participation being a relevant mediator. Hence, we treat this topic more systematically in this paper and split up the (total) effect of education on cognitive abilities into a direct effect and an indirect one going through the channel of labor-force participation.

In our analysis, we pool data from SHARE and ELSA on 76,000 observations from several countries in Europe across the years 2002–2017. The data include experimentally collected measures of cognitive abilities (including our main outcome, a word recall test as a proxy for fluid intelligence). We use compulsory schooling reforms and early retirement regulations as sources of exogenous variation and can replicate the effect of education on cognitive abilities as found in the literature before. On average, compliers to the compulsory-schooling reforms have a larger memory capacity: they recall 7.8 percent more words (24% of a SD) at age 50-70. Likewise, they are 18 percentage points less likely to be retired on average. We then examine how both effects interact with each other. As a result of our causal mediation analysis we find about a third of the total effect of education to run through labor force participation. This key finding of our paper

[4]Several other recent methodological advances make (sequential) conditional independence assumptions, as documented in Pearl (2001), Imai et al. (2010), Hong (2010), Tchetgen and Shpitser (2012) or Huber (2014).

may help to resolve the results in previous studies that the short-run effect of education on abilities seems to be small (or absent) while the long-run effect is often estimated to be large. It appears that not education directly, but how it affects choices later in life may drive the heterogeneous effects of education on cognitive abilities from prior studies. The additional finding that retirement does not cause a cognitive decline among the less educated compulsory schooling compliers singles out an important complementarity between education and old-age labor force participation. Because of these long-run implications of education, education itself may become an even more important target for policy.

This paper proceeds as follows. In Section 2 we describe the empirical approach and how we extend MTE estimation to causal mediation analysis. In Section 3 we demonstrate the validity of our approach by a simulation study. Section 4 describes the data. Estimation results are reported in Section 5 while Section 6 concludes.

2 Empirical approach: mediation analysis

2.1 Notation and parameters of interest

We start with the traditional potential outcome model, where Y^1 and Y^0 are the potential outcomes—cognitive abilities—with and without treatment. The binary treatment D is "more" education. The observed outcome Y either equals Y^1 in case an individual received a treatment or Y^0 in the absence of the treatment (we suppress the individual identifier i). We want to find out what drives the effect of D on the outcome variable Y. Technically, we want to decompose the total treatment effect ($E(Y^1 - Y^0)$ for the overall population or a certain subgroup) into components that are caused by potential mediating factors and a remaining part that may be the direct effect of the treatment.

In this paper, we focus on one mediating factor—the binary mediator variable M—but the setting can potentially be extended to accommodate any number of those factors. This mediator—retirement—may affect Y but may itself also be affected by D. It, likewise, is the realization of one of the two potential outcomes M^1 and M^0, where M^1 is labor force status with $D = 1$ and M^0 is labor force status with $D = 0$. To accommodate the analysis of direct and indirect treatment effects in a potential outcome model, we need to extend the notation to allow for hypothetical variations in the mediator M. The potential outcome then reads Y^{jk} where $j \in \{0, 1\}$ denotes the educational choice D and $k \in \{0, 1\}$ denotes the retirement choice M. For instance, Y^{11} is the potential outcome if the treatment and the mediator are chosen. We can also evaluate the potential outcome by fixing the treatment at j and letting the individual decide about the mediator depending on the hypothetical treatment state $l \in \{0,1\}$: Y^{jM^l}. Thus, j and l both refer to a potential treatment state of D. If $j = l$, individuals are assigned to a treatment state and then take their natural mediator choice under the same treatment state. If $j \neq l$, the quantity Y^{jM^l} informs about the outcome with treatment j if the individual chose the mediator as if the treatment choice was l. Note that in this case, this outcome is a counterfactual for every individual. The following hypothetical observation rule informs about how the potential outcome Y^{jk} translate into mediated outcomes Y^{jM^l}:

$$Y^{jM^l} = Y^{j1}M^l + Y^{j0}(1 - M^l)$$

It is a hypothetical observation rule if $j \neq l$. One can now define (at least) three different treatment effects that may be of interest, see, e.g. Frölich and Huber (2017). The first is the total treatment effect (TTE):

$$\text{TTE} = E\big(Y^{1M^1} - Y^{0M^0}\big)$$

Note that this notation is equivalent to the notation $E(Y^1 - Y^0)$, where any mediating factor is implicit.

The total treatment effect has two components: the direct effect of the treatment on the outcome and the indirect effect that goes through the mediator. The direct effect originates in the j part of the potential outcome written above, while the indirect effect arises due to differences in the k part. One can show this formally by decomposing the TTE into a direct component and an indirect one that runs through the mediator (see Huber, 2020). The direct treatment effect holds the mediator constant at either M^0 or M^1 and is therefore defined as

$$\text{DTE}(k) = E\big(Y^{1M^k} - Y^{0M^k}\big)$$

where $k \in \{0,1\}$. Thus, it closes down the effect D has on Y via M. The indirect treatment effect is defined vice versa. It fixes the treatment at $D = j$ and only varies the mediator from M^0 to M^1:

$$\text{ITE}(j) = E\big(Y^{jM^1} - Y^{jM^0}\big)$$

Adding and subtracting Y^{0M^1} or Y^{1M^0} from the TTE yields

$$\begin{aligned}
\text{TTE} &= E\big(Y^{1M^1} - Y^{0M^0}\big) \\
&= E\big(Y^{1M^1} - Y^{0M^1} + Y^{0M^1} - Y^{0M^0}\big) \\
&= E\big(Y^{1M^1} - Y^{0M^1}\big) + E\big(Y^{0M^1} - Y^{0M^0}\big) = \text{DTE}(1) + \text{ITE}(0) \\
&= E\big(Y^{1M^0} - Y^{0M^0}\big) + E\big(Y^{1M^1} - Y^{1M^0}\big) = \text{DTE}(0) + \text{ITE}(1) \qquad (1)
\end{aligned}$$

Differences between TTE and DTE(0) or DTE(1) arise only if two conditions hold jointly. First, D needs to change the individuals' choice behavior of M, i.e. $M^1 \neq M^0$ for a relevant subgroup of individuals. Second, these changes in the choice behavior need to induce meaningful changes in the k-part of the potential outcomes, i.e. $Y^{jM^1} - Y^{jM^0} \neq 0$ for the same group of individuals. We can calculate these treatment parameters by estimating the four magnitudes $E(Y^{1M^1})$, $E(Y^{1M^0})$, $E(Y^{0M^1})$, and $E(Y^{0M^0})$. Estimation of direct and indirect treatment effects will only be straightforward if we have two independent random assignments of D and M and full compliance. While being advantageous in terms of identification, such a setting may not be the most economically insightful in terms of the identified mediation effects, because in such a setting the researcher (deliberately or not) controls the size of an indirect effect. It will turn out that it is exactly the additional information that partial non-compliance carries that makes indirect treatment effects an insightful economic parameter.

2.2 The geometry of direct and indirect treatment effects

Mediation analysis is about how the two potential outcomes of M (M^1 and M^0) relate to the four potential outcomes of Y ($Y^{11}, Y^{10}, Y^{01}, Y^{00}$). We can visualize all aspects of our mediation analysis in a graph that plots conditional expectations of $Y^{j,k}$ against conditional expectations of M^j. We now describe this graphical approach, which we believe allows to gain more intuition about the causes of direct and indirect treatment effects, their necessary assumptions for identification as well as how to estimate them transparently. For simplicity, we first discuss a mediation analysis using a randomized controlled trial with perfect compliance. After this, we extend the setting to allow for the more interesting case of (partial) noncompliance. In doing so, think first about one hypothetical experiment, in which (with perfect compliance) we randomly assign D. As is well known, randomly assigning D directly allows for estimation of $E(Y^1)$ and $E(Y^0)$. Next, think about a second random assignment of the mediator M (with full compliance as well).

In Figure 2 we plot (conditional expectations of) potential outcomes of Y on the vertical axis for all individuals in the population. The population has a mass of one. If all individuals are sorted by some variable, the horizontal axis reports the quantiles of the distribution of this variable. Put differently, it shows, at each point, the cumulative share of individuals up to a certain rank. Here, we sort individuals by a crude measure: everybody who is assigned $M = 1$ in the experiment is on the left, while everybody who is assigned $M = 0$ is on the right.[5] Because M is random and, thus, independent of D, we have that $E(M^1) = E(M^0)$, since $P(M = 1|D = 1) = P(M = 1|D = 0)$. Moreover, since D and M are random, they are independent of Y^{jk}, hence, the four potential outcome lines (the red dashed lines in the graph) are flat and without any discontinuities. Those in the left part with $M = 1$ have the same potential outcomes as those in the right part with $M = 0$. The blue solid lines are the resulting $E(Y^1)$ and $E(Y^0)$, as shown in the observation rule in the graph. At $E(M^1)$ and $E(M^0)$ (essentially the share of individuals assigned $M = 1$) there is a discontinuous jump where M switches from 1 to 0 and $E(Y^1)$ from $E(Y^{11})$ to $E(Y^{10})$ (and likewise for $E(Y^0)$). The area between the blue solid lines is the average total treatment effect in the population.

Next, in Figure 3, we consider a case where the assignments of D and M are still random, but where the assignment probability of $M = 1$ is higher if $D = 1$ as compared to $D = 0$. This induces an effect of D on M. In this case, we still have—due to randomization—four potential outcome lines that are unrelated to the assignment of M (and, thus, are flat). However, as Figure 3 shows, the two resulting lines $E(Y^1)$ and $E(Y^0)$ do not exhibit the discontinuous jump at the same value of the horizontal axis. $E(Y^1)$ jumps at $E(M^1)$ while $E(Y^0)$ jumps at $E(M^0)$. This gives rise to indirect treatment effects, which measure the interdependence between the assignment of D and M in terms of the corresponding effect on Y. The magnitude of this interdependence is easy to quantify and consists of two separate components that result in the blue-shaded area in Figure 3. First, the jump at $E(M^1)$ informs about the causal effect of M on $E(Y^1)$ (which is A-B in the figure). Second, the effect of D on M is the line C-B. Thus, the ITE(1) is the causal effect of M on $E(Y^1)$ multiplied by the causal effect of D on M.

The geometry of these indirect treatment effects stresses that either one of its components (effect of M on $E(Y^j)$ or the effect of D on M) is only a necessary but not a sufficient condition for M

[5] At this point, it does not matter how individuals are sorted within the two strata of M. This will change below.

Figure 2: Potential outcomes in a randomized controlled trial

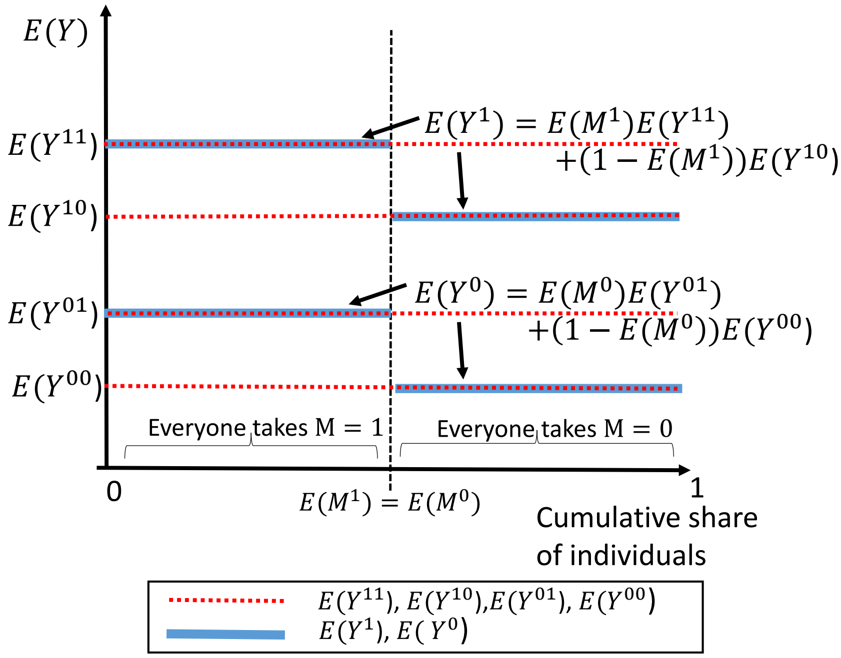
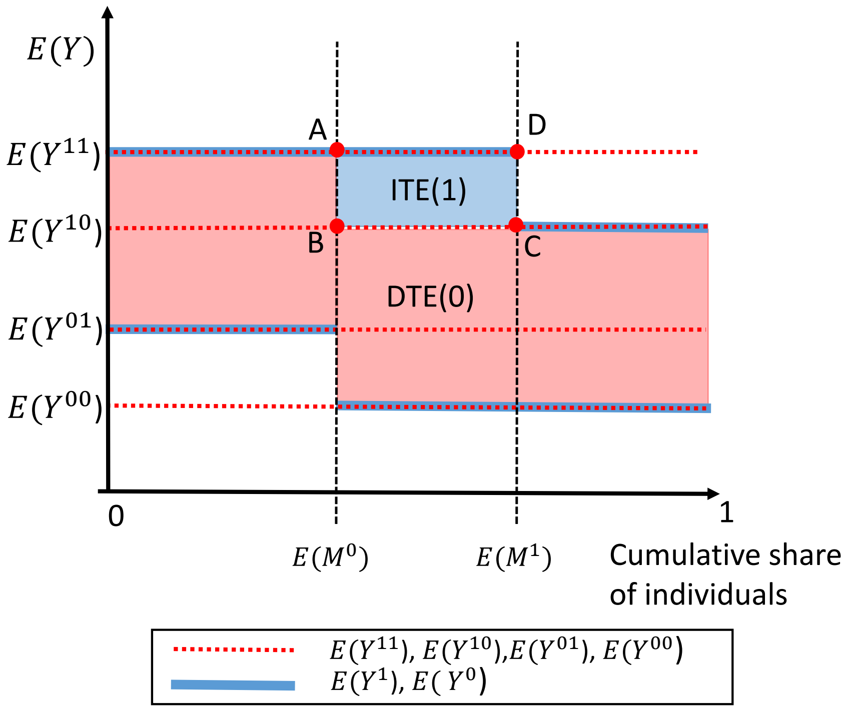


Figure 3: Potential outcomes in a randomized controlled trial with effect of D on M



being a mechanism of the total treatment effect—they need to hold jointly. In Figure 2, there is no effect of D on M and, hence, no ITE. This is an important observation for many applied papers that assess the role of potential mechanisms by only estimating the effect of a certain treatment of a particular mediator. In our notation, these papers estimate $M^1 - M^0$ only. Independent of the size of this effect, however, the ITEs will be zero if M influences neither Y^1 nor Y^0. The TTE still is the area between the solid lines but can now be split up into an ITE and a DTE (which is the red-shaded area).

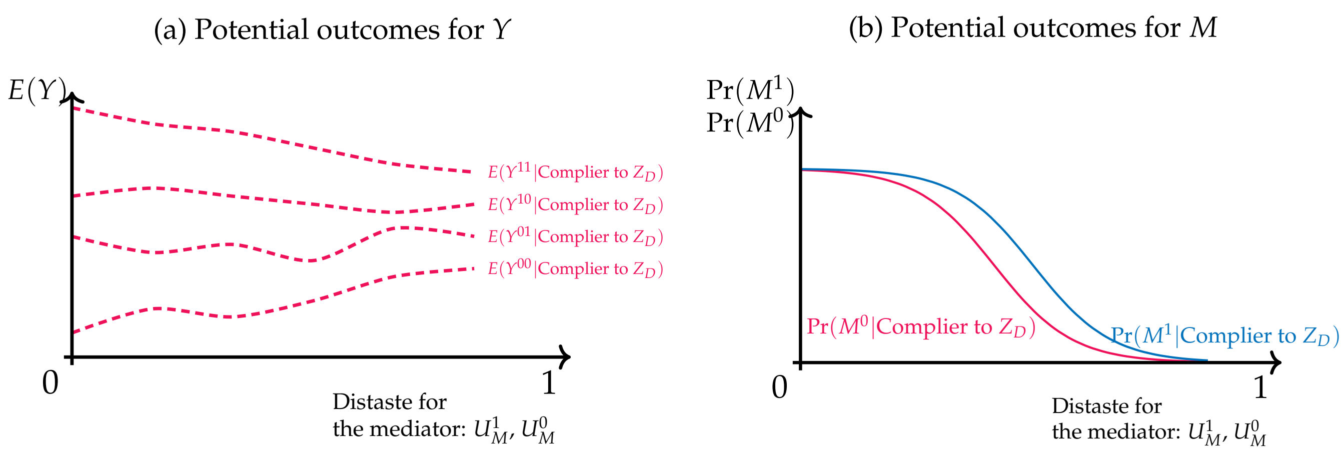
Figure 3 also shows the thought experiments that are necessary to determine ITE and DTE. For the ITE(1), we compare the realized Y^1 (which is Y^{1M^1}) with a hypothetical Y^1 that is determined by using M^0, that is, we move the blue solid line from A-D to B-C. For DTE(0), we compare Y^1 and Y^0 while fixing M at M^0, that is, again moving the blue solid line from A-D to B-C. As in Equation (1), both ITE(1) and DTE(0) add up to the TTE. Figure A1 in the Appendix shows the equivalent case of ITE(0) and DTE(1).

2.3 Direct and indirect treatment effects with non-compliance and heterogeneous treatment effects

Next, assume that D still is randomized with full compliance but we allow for non-compliance/self-selection of M. Additionally, we explicitly allow for heterogeneous treatment effects (which we implicitly allowed for before but with non-compliance, heterogeneity in individual treatment effects becomes irrelevant as the random assignment averages it out). We can think of the new assignment mechanism as instrument Z_M that randomly assigns some sort of incentive to take M. Still, the empirical task is to estimate the four potential outcome and the two potential mediator curves, $E(Y^{jk})$ and $E(M^j)$, respectively, in order to form the treatment parameters we are interested in. However, self-selection implies that $E(M^j)$ are no longer vertical lines, where on the left, everyone takes $M = 1$ and on the right, no one does. Instead, $E(M^j)$ are monotonically declining functions along the horizontal axis. This is because with non-compliance, we can only order individuals according to some measure of the hypothetical incentive necessary to make individuals take $M = 1$. This measure is called distaste to take the treatment (in our case the mediator) in the marginal treatment effects literature (see, e.g. Heckman and Vytlacil, 2005) and is also normalized to the unit interval. The results could be plotted as in Figure 4 and, in what follows, we describe all its ingredients. Panel 4a, which already also captures potential endogeneity of D, treated below, depicts the potential outcomes of Y, whereas Figure 4b plots the potential outcomes of M.

Since D is random, we can—abstracting from covariates in case of conditional randomization—easily estimate $E(Y^1)$ as sample mean of observed Y among all treated and $E(Y^0)$ as sample mean of observed Y among all untreated. Thus, only the k-part of $E(Y^{jk})$ poses problems. To solve this, we use the insights and methods of Carneiro and Lee (2009). They show how to use instrumental variables to estimate conditional means of potential outcomes Y^1 and Y^0 of one endogenous treatment along this distaste for the treatment along which individuals are indifferent between taking and not taking the treatment. Mogstad et al. (2018) call these marginal treatment response (MTR) functions. The difference between MTRs of the treated and untreated outcomes are MTEs. Transferred to our problem, we can exploit Z_M to estimate conditional expectations of Y^{11} and Y^{10} along the distribution of the unobserved heterogeneity in M for the treated (with respect to D) and the corresponding quantities for Y^{01} and Y^{00} for the untreated. They are estimated and plotted for each value of distaste to take mediator M which, in accordance with the literature, could be named U_M and be normalized to the unit interval. U_M sorts individuals in a similar way on the horizontal axis as we did in Figures 2 and 3. The major difference is that we only had two polar cases before: after randomization there was no choice left and $E(M) = 1$ for everybody on the left, and $E(M) = 0$ for everybody on the right. Sorting by U_M allows for continuous mediator-choice probabilities between 0 and 1. Still, the horizontal axis sorts individuals according to their choice-probability (or the revealed preference for the mediator) as before. Because conditional on U_M, M is randomized (and conditional on compliers to Z_D, D is also random, both with perfect compliance), we can identify counterfactual mediator (and treatment) choices in the outcome.

Figure 4: Mediation analysis with non-compliance and heterogeneous treatment effects



We derive this more formally now and use a generalized Roy model in which an individual weights their cost and benefits and will choose $M^j = 1$ if the net benefit is positive (e.g., see

Heckman, 2010). Recall that the decision to choose M may potentially depend on the initial assignment of D. We allow for this by adding the superscript j. Let $Y^{j1} = \mu^{j1} + U^{j1}$ and $Y^{j0} = \mu^{j0} + U^{j0}$ be the potential outcome functions that can be separated into functions of observables (μ^{j1} and μ^{j0}) and unobservables (U^{j1} and U^{j0}). Choice of M induces costs that can, likewise, be separated into an observable and an unobservable part: $C^j(Z_M) = \mu^j_C(Z_M) + v^j$. The instrument Z_M does not affect the potential outcomes but the cost-function and, via this, the choice of M. Individuals choose $M = 1$ if expected benefits outweigh expected costs:

$$
\begin{aligned}
M^j &= \mathbb{1}\Big[Y^{j1} - Y^{j0} \geq C^j(Z_M)\Big] \\
&= \mathbb{1}\Big[\mu^{j1} - \mu^{j0} - \mu^j_C(Z_M) \geq U^{j1} - U^{j0} - v^j\Big] \\
&= \mathbb{1}\Big[\mu^j_M(Z_M) \geq \mathcal{V}^j\Big] \\
&= \mathbb{1}\Big[F_{\mathcal{V}}\Big(\mu^j_M(Z_M)\Big) \geq F_{\mathcal{V}}\Big(\mathcal{V}^j\Big)\Big] \\
&= \mathbb{1}\Big[\Pr(M = 1|Z_M) \geq U^j_M\Big] \qquad \forall j \in \{0, 1\}.
\end{aligned} \tag{2}
$$

The second step collects observed components on the left, and unobserved components on the right side of the inequality, respectively. Subsequently, the inequality is simplified by renaming $\mu^j_M(Z_M) = \mu^{j1} - \mu^{j0} - \mu^j_C(Z_M)$ for the observed and $\mathcal{V}^j = U^{j1} - U^{j0} - v^j$ for the unobserved components. Finally, we monotonically transform both sides of the inequality by $F_{\mathcal{V}}(\cdot)$. This transformation yields the ranks (U^j_M) when applied to $\mathcal{V}$ and the probability to retire when applied to the index of observed factors (μ^j_M). This probability is referred to as the propensity score for any binary treatment. By definition, U^j_M ranges between 0 and 1 and represents the respective threshold value for an individual with a given set of observable factors (that determine the propensity score) to select into the treatment, here $M = 1$. At a specific threshold $p \in (0, 1)$, where $U^j_M = p$, individuals are indifferent between taking the treatment or not and reveal their taste or distaste for M by opting for it at a specific value of the propensity score. Applying these principles to individuals who have a propensity score of close to zero, we can infer their value of U^j_M for individuals who nonetheless take $M = 1$. It has to be even smaller than the propensity score for everyone and hence U^j_M measures an inverse score of the revealed preferences.[6]

[6]Also note that our approach can allow U^j_M to differ by treatment state: $U^1_M \neq U^0_M$. Hence, we do not need to impose any rank invariance restriction for the mediation analysis. Of course, education may change the preference ordering for retirement. Nevertheless, this would never imply direct and indirect effect on the aggregate U_M level, if the relevant

Note that the last step of Eq. (2) shows how to estimate the two quantities $E(M^j|U_M^j)$ plotted in Figure 4b. For this, we just need to determine the fraction of estimated propensity score values larger than the respective U_M^j value at which we evaluate: $E\big(\mathbb{1}[\Pr(M=1|Z_M) \geq U_M^j]|U_M^j\big)$. By definition, $E(M^j|U_M^j)$ has to decay along U_M^j. In the MTE literature, these quantities are known as the weights for aggregating the MTEs to a single ATT parameter. For our mediation analysis, these curves are weights for aggregating the Y^{j1} and Y^{j0} quantities into the relevant Y^{jM^0} and Y^{jM^1} outcomes. As before, a difference between $E(M^1|U_M^1)$ and $E(M^0|U_M^0)$ is a necessary condition for indirect treatment effects.

The index of the unobserved factors plays a crucial role in gaining more insights on how the selection into a treatment is correlated with the effects it evokes. We define

$$\text{MTR}^{jk}(p) \quad = \quad E(Y^{jk} \mid U_M = p)$$

where MTR stands for *marginal treatment response function*. We show below how we estimate this and how instruments are key for its identification.

Finally, we allow D to be a choice variable, too, hence we allow for a second dimension of endogeneity. We solve this by transferring the insights of Imbens and Rubin (1997) to our setting. Using a binary instrument Z_D, Imbens and Rubin (1997) show how to estimate $E\left(Y^1|\text{Complier to } Z_D\right)$ and $E\left(Y^0|\text{Complier to } Z_D\right)$ separately. We provide details in the next subchapter.

2.4 Estimation and identifying assumptions

Estimation of the MTRs and identifying assumptions

We believe that it is instructive to, again, start with a completely random and full-compliance D when we discuss the estimation of $\text{MTR}_C^{jk}(p)$ and extend to the case of non-compliance later. We start with the estimation of $\text{MTR}^{11}(p)$ and $\text{MTR}^{10}(p)$.

Assumption 1: We have an instrument Z_M that fulfills the typical LATE assumptions: strength with respect to M, conditional independence of the potential outcomes Y^{jk}, exclusion restriction, and monotonicity ("no defiers").

curves in Figure 4 are left unchanged by this. We document this in the simulation exercise of Section 3, where we restrict U_M^1 and U_M^0 to be uncorrelated, thereby ensuring that ranks may differ by treatment state.

Transferred to our setting, Carneiro and Lee (2009) and Brinch et al. (2017) (in their "separate estimation approach" of the marginal treatment effect) show that with such an instrument we can express the MTR at a specific value of the propensity score $P(Z_M) = Pr(M = 1|Z_M, \boldsymbol{X}) = p$ as

$$\begin{aligned} MTR^{11}(\boldsymbol{X} = \boldsymbol{x}, U_M = p) &= E(Y|\boldsymbol{X} = \boldsymbol{x}, P(Z_M) = p, D = 1, M = 1) \\ &\quad + p\frac{\partial E(Y|\boldsymbol{X} = \boldsymbol{x}, P(Z_M) = p, D = 1, M = 1)}{\partial p} \end{aligned} \tag{3}$$

$$\begin{aligned} MTR^{10}(\boldsymbol{X} = \boldsymbol{x}, U_M = p) &= E(Y|\boldsymbol{X} = \boldsymbol{x}, P(Z_M) = p, D = 1, M = 0) \\ &\quad -(1-p)\frac{\partial E(Y|\boldsymbol{X} = \boldsymbol{x}, P(Z_M) = p, D = 1, M = 0)}{\partial p} \end{aligned} \tag{4}$$

Note that, while we left control variables implicit before, we now make this explicit by conditioning on the vector of observables $\boldsymbol{X}$. These $\boldsymbol{X}$ are used in the propensity score estimation but, for simplicity, we write $P(Z_M)$ only. The expectation of Y, conditional on $\boldsymbol{X}$, p, and the mediator status $M = 1$ can be expressed as

$$E(Y|\boldsymbol{X} = \boldsymbol{x}, P(Z_M) = p, M = 1, D = 1) = \mu^{11}(\boldsymbol{x}) + f^{11}(p), \tag{5}$$

where $f^{11}(p)$ is an unknown function of p. Likewise for $M = 0$

$$E(Y|\boldsymbol{X} = \boldsymbol{x}, P(Z_M) = p, M = 0, D = 1) = \mu^{10}(\boldsymbol{x}) + f^{10}(p). \tag{6}$$

Estimation of the MTR is, then, straightforward. We stratify the sample by D and M, specify $\mu^{jk}(\boldsymbol{X})$ as a linear function of $\boldsymbol{X}$ plus linear interactions with p captured by the parameter vectors $\boldsymbol{\gamma^{jk}}$ and $\boldsymbol{\delta^{jk}}$, respectively. Hence, we estimate the parameters in the regression $Y = \boldsymbol{X'}\boldsymbol{\gamma}^{\boldsymbol{jk}} + \boldsymbol{X'}p\boldsymbol{\delta}^{\boldsymbol{jk}} + f^{jk}(p) + \varepsilon$ and calculate the MTR according to Eq. (3) and Eq. (4). Estimation could either be semi-parametric as, e.g. in Westphal et al. (2022) where no functional form of $f^{jk}(p)$ is assumed and the estimator suggested by Robinson (1988) is employed. Alternatively, the shape of $f^{jk}(p)$ is approximated by the inclusion of polynomials of p. The MTR^{0k} are estimated equivalently using the subsample of $D = 0$.

Since we have a binary instrument, the propensity score will only take on two values, conditional on $\boldsymbol{X}$. This means that we also need variation in $\boldsymbol{X}$ to receive a continuous p and to identify the MTR without making the strong assumption that the MTR is linear, see Brinch et al. (2017). This can be achieved by an additive separability assumption. While this assumption is already included in our specification of potential outcomes above, we spell it out here for completeness.

Assumption 2: Additive separability between observed and unobserved heterogeneity in the MTR. This means that $E(Y^{jk}|U_M, \boldsymbol{X} = \boldsymbol{x}) = \mu^{jk}(\boldsymbol{x}) + E(U^{jk}|U_M)$.

This assumption restricts the slope (but not the level) of the MTR to be equal across different $\boldsymbol{X}$ cells, i.e. treatment effects can vary with X and U_M, but not along their interaction (Brinch et al., 2017). With this assumption, we can additionally exploit variation in $\boldsymbol{X}$ for the full support condition and to identify local effects at different values of $P(Z_M)$ for compliers to Z_M. Note that this assumption is standard in applied work, for instance, the conventional two-stage least squares approach that does not interact the treatment D with $\boldsymbol{X}$ implicitly makes an even stronger assumption (that treatment effects are additively separable in D and X) without further discussion. Moreover, MTEs are generally not only derived with this implicit separability assumption, they are also estimated with it. Without it, one would need to separately estimate MTEs for all possible combinations of $\boldsymbol{X}$, which quickly becomes infeasible. Examples of MTE applications exploiting additive separability include, for instance, Carneiro et al. (2011); Brinch et al. (2017); Nybom (2017); Cornelissen et al. (2018).

Next, as our treatment, education, is endogenous, we allow for self-selection into D and adapt the approach of Imbens and Rubin (1997) to our setting.

Assumption 3: We have an instrument Z_D that fulfills the typical LATE assumptions: strength with respect to D, conditional independence of the potential outcomes Y^{jk}, exclusion restriction, and monotonicity ("no defiers").

We first show how to estimate the total treatment effect.[7] We start by estimating the share of Z_D-compliers by the following first-stage regression of D on the instrument Z_D and a set of control variables, possibly to justify the conditional independence assumption of the instrument. Assume that, throughout this exposition, all control variables enter as mean-centered values so that $E(X) = 0$.

$$D = \pi_0 + \pi_1 Z_D + \boldsymbol{X'}\boldsymbol{\pi}^{\boldsymbol{X}} + u \tag{7}$$

Then, by monotonicity, π_0 gives the share of always-takers (AT), as $E(D|X, Z_D = 0) = \pi_0$. The share of compliers (C) is π_1, reflecting individuals who change D when Z_D switches from 0 to 1. The share of never-takers (NT), thus, is $1 - \pi_0 - \pi_1$.

[7] Of course, this can simply be estimated by two-stage least squares and our approach delivers the algebraically same results.

Knowing these shares, we can estimate Y^1 and Y^0 for the compliers, to get the (local average) total treatment effect according to Imbens and Rubin (1997). This is possible with the following regression:

$$
\begin{aligned}
Y &= \delta_0 + \delta_1 \Big[\mathbb{1}(D=1)\mathbb{1}(Z_D=0)\Big] + \delta_2 \Big[\mathbb{1}(D=1)\mathbb{1}(Z_D=1)\Big] \\
&\quad + \delta_3 \Big[\mathbb{1}(D=0)\mathbb{1}(Z_D=0)\Big] + \delta_4 \Big[\mathbb{1}(D=0)\mathbb{1}(Z_D=1)\Big] + \boldsymbol{X'\delta^X} + v \qquad (8)
\end{aligned}
$$

The coefficient δ_1 gives the always-taker-specific mean of Y (if the $\boldsymbol{X}$ variables are not demeaned, the coefficients need to be adjusted). Assumption 3 ensures that $\delta_1 = E(Y^1|AT)$. Likewise, δ_2 gives the mean for a subgroup formed of always-takers and (treated) compliers (that is, individuals with $D = 1$ and $Z_D = 1$). Vice versa, δ_4 is the mean of never-takers (thus, $\delta_4 = E(Y^0|NT)$), whereas δ_3 is the mean of untreated compliers and never-takers. To estimate expected outcomes for treated and untreated compliers, we use δ_2 and δ_3 and adjust them for δ_1 and δ_4, with the help of the group shares. Thus,

$$
\begin{aligned}
E(Y^1|C) &= \frac{\delta_2(\pi_0+\pi_1) - \delta_1\pi_0}{\pi_1} \\
E(Y^0|C) &= \frac{\delta_3(1-\pi_0) - \delta_4(1-\pi_0-\pi_1)}{\pi_1}.
\end{aligned}
$$

The TTE is the difference between these quantities. Next, we do the same to get the $E(Y^{jk}|C)$ and jointly address endogeneity in D and M. This means to run Eq. (5) and Eq. (6) each four times for all possible strata of D, Z_D and determine MTR^{11}_{AT}, $\text{MTR}^{11}_{AT,CT}$,MTR^{10}_{AT}, $\text{MTR}^{10}_{AT,CT}$,MTR^{01}_{NT}, $\text{MTR}^{01}_{NT,CU}$,MTR^{00}_{NT}, and $\text{MTR}^{00}_{NT,CU}$. Here, as before, AT stands for always-takers of D and NT for never-takers of D. CT is treated compliers (compliers to Z_D and $Z_D = 1$) and CU is untreated compliers (compliers to Z_D and $Z_D = 0$). $\text{MTR}^{11}_{AT,CT}$ is the MTR identified for the group of always-takers and compliers and we follow the same approach as above to derive MTR^{11}_{CT}.

To ensure that assumption 1 (randomization of Z_M) holds even when conditioning on Z_D, we need to make an additional assumption:

Assumption 4: Conditional independence of the two instruments Z_M and Z_D.

This is necessary when stratifying the group of Z_D compliers further by their reaction to Z_M. We test one implication of this assumption later, by evaluating whether the instrument Z_D gives us any predictive power to determine Z_M.

We could either purge the influence of the observable characteristics $\boldsymbol{X}$ from the dependent variable by OLS or the estimator proposed by Robinson (1988). We opt for the former for convenience[8] and non-parametrically regress this cleaned variable on the propensity score at different evaluation points over the unit interval of the propensity score. With the local slope and the level coefficients at specific values of the propensity score over the unit interval, we can apply Eq. (5) and Eq. (6) to compute the eight MTRs formed by every combination of Z_D, D, and M.

Again, holding fixed D and M, differences between the curve with $Z_D = 1$ and the corresponding one with $Z_D = 0$ must arise only because of the complier if the typical LATE/IV assumptions hold. Because we know the share of compliers, we can to apply the principle of Imbens and Rubin (1997) and calculate

$$
\begin{aligned}
MTR_C^{1k}(U_M = p) &= \frac{MTR_{AT,CT}^{1k}(U_M = p)(\pi_0 + \pi_1) - MTR_{AT}^{1k}(U_M = p)\pi_0}{\pi_1} \\
MTR_C^{0k}(U_M = p) &= \frac{MTR_{NT,CU}^{0k}(U_M = p)(1 - \pi_0) - MTR_{NT}^{0k}(U_M = p)(1 - \pi_0 - \pi_1)}{\pi_1} \\
& \qquad \forall k \in \{0, 1\}.
\end{aligned}
$$

Finally, we also make the following assumption, which is necessary to identify direct and indirect effects for the population of Z_D-compliers (and not just a subset, e.g. defined by Z_M compliers).

Assumption 5: Full support of PS_M by $M = 1$ and $M = 0$ within the group of compliers of D.

[8] To do this, we run, as a pre-processing step, the following regression where we restrict the control variables to have the same effect for all subgroups (the results are robust with respect to the chosen pre-processing method):

$$
\begin{aligned}
Y &= \alpha_{DZ_DM} + \beta_{DZ_DM}p + \boldsymbol{X'}\boldsymbol{\gamma} + \boldsymbol{X'}p\boldsymbol{\delta} + \varepsilon \\
\tilde{Y} &= Y - (\boldsymbol{X} - \overline{\mathbf{X}})'\widehat{\boldsymbol{\gamma}} - (\boldsymbol{X} - \overline{\mathbf{X}})'p\widehat{\boldsymbol{\delta}}
\end{aligned}
$$

Then, we compute an adjusted dependent variable $\tilde{Y}$ in the equation above where the $\boldsymbol{X}$ values are fixed at the respective mean values and using the estimated coefficients from this regression. The α_{DZ_DM} and β_{DZ_DM} coefficients are specific levels and slopes for the propensity score, respectively and may differ between the eight possible combinations of D, M, and Z_D. This ensures that $\tilde{Y}$ contains all marginal effects of M on Y in each $D \times Z_D$ cell (which inform about the corresponding Z_D types) but without $\boldsymbol{X}$ confounding these correlations. Hence, we can use $\tilde{Y}$ as the dependent variable in eight non-parametric local linear regressions on the propensity score p.

Estimation of $\Pr(M^1)$, $\Pr(M^1)$ and mediated outcomes Y^{1M^1}, Y^{1M^1}, Y^{1M^1}, and Y^{1M^1}

Thus, we know how to estimate the required four potential outcome curves for our mediation approach. The last missing pieces are $E(M^1|U_M, C)$ and $E(M^0|U_M, C)$ for the Z_D compliers denoted by C and the mediated outcome curves (estimates of Y^{jM^l}) that result from it. Regarding the former, we can simply compute $E\big(\mathbb{1}[\Pr(M = 1|Z_M) \geq U^j_M]|U^j_M\big)$ along U_M in the four cells determined by D and Z_D, as outlined above, and, in the final step, apply the simple Imbens and Rubin (1997) formula to adjust the quantity to the compliers.

Now, we can identify the six quantities MTR^{jk}_C and $E(M^j|U_M, C)$, which are conditional expectations Y^{jk} and M^j, respectively, for the compliers along U_M. In a final step, we need to compute mediated outcomes Y^{jM^l}, as these determine the total, direct, and indirect effects. This is straightforward, as when applying $E(\cdot|U_M, C)$ to the hypothetical observation rule $Y^{jM^l} = Y^{j1}M^l + Y^{j0}(1 - M^l)$, it shows that $E(Y^{jM^l}|U_M, C)$ is a weighted mean of MTR^{j1}_C and MTR^{j0}_C, with weights determined by $E(M^l|U_M, C)$. Because the MTRs and $E(M^j|U_M)$ are average outcomes along the margin of indifference, we can solve the identification problem inherent in Y^{1M^0} and Y^{0M^1}, which are never observed. This is because at each value of U_M, compliers of D (our target group) are similar in terms of their observable and unobservable characteristics (and D and M are randomized with perfect compliance). Hence, we know their outcome if they had chosen a different treatment and mediator state and can also disentangle the outcome effect from the mediator effect at this margin. This is why the MTE approach is so appealing for mediation anlyses.

In our setting, instrument Z_D is compulsory schooling. Compliers are individuals who have to extend their years in school because of the enforcement of a new compulsory schooling threshold. Because this minimum amount of years of schooling is mandatory for everyone, defiers (individuals who drop out of school and get fewer years of schooling because of the reform) and never-takers (individuals who drop out of school before attaining the compulsory amount of years irrespective of the reform) cannot exist.

Standard errors of all effects are estimated by bootstrapping the entire procedure for 200 times.

3 Simulation studies

We carry out three different simulation studies to compare the performance of our estimator with a classic instrumental variables approach. Data generating process 1 (**DGP 1**) is the following simple case without control variables and a continuous instrument Z_2:

$$
\begin{aligned}
Y^{11} &= \alpha_D + \alpha_M + \alpha_{DM} + U^{11} \\
Y^{10} &= \alpha_D \qquad\qquad\quad + U^{10} \\
Y^{01} &= \quad + \alpha_M \qquad\quad + U^{01} \\
Y^{00} &= \qquad\qquad\qquad\quad U^{00} \\
M^1 &= \mathbb{1}(\gamma_Z^1 Z_2 + \alpha + U_M^1 > 0) \\
M^0 &= \mathbb{1}(\gamma_Z^0 Z_2 \qquad + U_M^0 > 0) \\
D &= \mathbb{1}(\beta_Z^D Z_1 + V > 0) \\
Z_1 &= \mathbb{1}(P > 0) \\
Z_2 &= Q \\
M &= M^1 D + M^0 (1 - D) \\
Y &= DMY^{11} + D(1-M)Y^{10} + (1-D)MY^{01} + (1-D)(1-M)Y^{00}
\end{aligned}
$$

P, Q independent standard normal. The parameters are set to $\alpha_D = 1$, $\alpha_M = 1$, $\alpha_{DM} = 5$, $\alpha = 1$, $\beta_Z^D = 1$,$\gamma_Z^1 = 1$, $\gamma_Z^0 = 1$. Moreover, the error correlation structure is as follows:

$$
\begin{bmatrix} U^{11} \\ U^{10} \\ U^{01} \\ U^{00} \\ U_M^1 \\ U_M^0 \\ V \end{bmatrix} \sim \mathcal{N} \left[\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.32 & 0.32 & 0.32 & 0.32 & 0.32 & 0.32 \\ 0.32 & 1 & 0.32 & 0.32 & 0.32 & 0.32 & 0.32 \\ 0.32 & 0.32 & 1 & 0.32 & 0.32 & 0.32 & 0.32 \\ 0.32 & 0.32 & 0.32 & 1 & 0.32 & 0.32 & 0.32 \\ 0.32 & 0.32 & 0.32 & 0.32 & 1 & 0.32 & 0.32 \\ 0.32 & 0.32 & 0.32 & 0.32 & 0.32 & 1 & 0.32 \\ 0.32 & 0.32 & 0.32 & 0.32 & 0.32 & 0.32 & 1 \end{bmatrix} \right]
$$

All error terms are correlated and, hence, D and M are endogenous. DGP 1 generates a complier share of 34% and a correlation of U_D and U_M of 0.32. We vary the error term structure in two ways:

DGP 1′ (stronger correlation):

$$\begin{bmatrix} U^{11} \\ U^{10} \\ U^{01} \\ U^{00} \\ U_M^1 \\ U_M^0 \\ V \end{bmatrix} \sim \mathcal{N}\left[\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 & 0.8 & 0.8 & 0.8 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 & 0.8 & 0.8 & 0.8 & 0.8 \\ 0.8 & 0.8 & 1 & 0.8 & 0.8 & 0.8 & 0.8 \\ 0.8 & 0.8 & 0.8 & 1 & 0.8 & 0.8 & 0.8 \\ 0.8 & 0.8 & 0.8 & 0.8 & 1 & 0.8 & 0.8 \\ 0.8 & 0.8 & 0.8 & 0.8 & 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 0.8 & 0.8 & 0.8 & 0.8 & 1 \end{bmatrix}\right]$$

DGP 1″ (No correlation between U^{11}, U^{10}, U^{01}, and U^{00}, and no correlation between U_M^1 and U_M^0):

$$\begin{bmatrix} U^{11} \\ U^{10} \\ U^{01} \\ U^{00} \\ U_M^1 \\ U_M^0 \\ V \end{bmatrix} \sim \mathcal{N}\left[\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 & 0.32 & 0.32 & 0.32 \\ 0 & 1 & 0 & 0 & 0.32 & 0.32 & 0.32 \\ 0 & 0 & 1 & 0 & 0.32 & 0.32 & 0.32 \\ 0 & 0 & 0 & 1 & 0.32 & 0.32 & 0.32 \\ 0.32 & 0.32 & 0.32 & 0.32 & 1 & 0 & 0.32 \\ 0.32 & 0.32 & 0.32 & 0.32 & 0 & 1 & 0.32 \\ 0.32 & 0.32 & 0.32 & 0.32 & 0.32 & 0.32 & 1 \end{bmatrix}\right]$$

Estimation and results:

Figure 5 and Figures S1 and S2 as well as Table S1 in the Supplementary Materials report the results using 200 rounds of simulation with 50,000 observations per round. The two vertical lines in each subfigure denote the (unconditional) indirect and direct effects, both for the full sample as well as for the subsample of compliers to instrument Z_1. We aim at estimating the effects for the compliers and compare average estimated effects to the average true effects. The bars are averages over the 200 estimates gained by four different estimation procedures. While Figure 5 shows results for DGP 1, Figures S1 and S2 show results for DGP 1′ and DGP 1″, respectively.

As a benchmark, we carry out OLS regressions where the parameters of the following two equations are estimated:

$$
\begin{aligned}
Y &= \delta_0 + \delta_D D + \delta_M M + \delta_{DM} DM + \delta_X X + \epsilon \\
M &= \gamma_0 + \gamma_D D + \gamma_X X + u
\end{aligned}
$$

The mediation effects are estimated to be

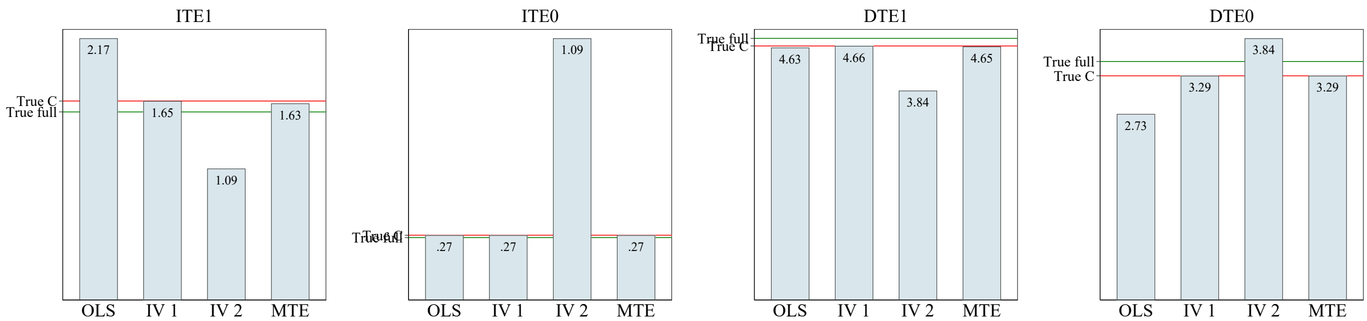
$$
\begin{aligned}
ITE(1) &= \gamma_D \cdot (\delta_M + \delta_{DM}) \\
ITE(0) &= \gamma_D \cdot \delta_M \\
DTE(1) &= \delta_D + \delta_{DM} \cdot (\gamma_0 + \gamma_D) \\
DTE(0) &= \delta_D + \delta_{DM} \cdot \gamma_0,
\end{aligned}
$$

and TTE (not reported in the figures but in the table) is derived using Eq. (1).

The results for the bars denoted by *IV 1* in Figure 5 are generated by the same procedure with the only difference that the two equations above are estimated by two-stage least squares. Here, Z_1 is used as an instrument for D, Z_2 is the instrument for M and the interaction of Z_1 and Z_2 is used as an additional instrument in the equation with the interaction of D and M. Following Frölich and Huber (2017), we also present results when possible interaction effects of D and M are ignored in the IV estimations as sometimes seen in the literature. This means to assume ITE(1) = ITE(0) and DTE(1) = DTE(0). The results are reported by the bars denoted *IV 2*. Finally, we report the results of our proposed estimator in the bars denoted by *MTE*.

The average effects (and, thus, biases when compared to the true effects) are reported in the figures while Table S1 in the Supplementary Materials also reports the root mean squared error in the columns denoted by RMSE. DGP 1 shows that *OLS* produces biased estimates, particularly in the estimation of ITE(1) and DTE(0). For instance, while the average true ITE(1) for compliers in DGP 1 is 1.65, OLS estimates it to be 2.17. The bias is aggravated when the errors are correlated more strongly as in DGP 1′ (see Figure S1). Likewise, a wrongly specified instrumental variables estimation (as in *IV 2*) yields estimates of ITE and DTE that are far away from the true effects. However, the more flexible IV specification with an interaction of D and M in *IV 1* only generates a negligible bias, where all average estimates are very close to the true average effects for the group of compliers. The same holds for our proposed estimator, as seen in line *MTE*. *IV 1* slightly dominates *MTE*, also in terms of RMSE, but, overall, the differences are negligible.

Figure 5: Results of DGP 1



Notes: Own calculations. This Figure plots the average results of DGP 1 after replicating the estimation respective procedures 200 times. Each panel refers to a specific indirect of direct mediation effect. The green and red horizontal line depict the average true effect for the whole sample and for the Z_1 compliers only (our target parameter), respectively. OLS and IV1 refer to implied mediation effects when $Y = \delta_0 + \delta_D D + \delta_M M + \delta_{DM} DM + \delta_X X + \epsilon$ and $M = \gamma_0 + \gamma_D D + \gamma_X X + u$ are estimated by OLS and two-stage least squares (2SLS), respectively. IV2 reports the implied effect when no interaction term is used in the 2SLS outcome regression. MTE refers to the results of the estimation procedure derived in the paper. Full results including RMSE are reported in Table S1.

The results for DGP 1′ and DGP 1″ are very similar and confirm that our method works also with different structures of error correlation.

DGP 2: Binary instruments and control variables

DGP 2 is meant to show that our approach works with two binary instruments (Z_1 and Z_2) and control variables. The DGP is in the spirit of the one proposed by Frölich and Huber (2017) but adapted to a binary instrument Z_2.

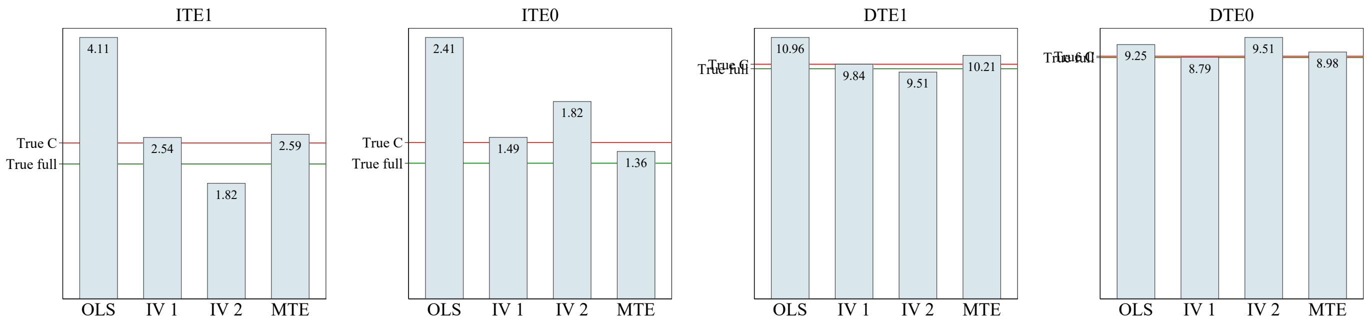
$$\begin{aligned} Y &= 5D + 10M + 7DM + 5X + U \\ M &= 1(-1.5 + 4Z_2 + 1D + 0.5X + V > 0) \\ D &= 1(-1.5 + 2Z_1 + 0.5X + W > 0) \\ Z_1 &= 1(0.5X + P > 0) \\ Z_2 &= 1(Q > 0) \end{aligned}$$

and

$$\begin{bmatrix} U \\ V \\ W \end{bmatrix} \sim \mathcal{N}\left[\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.6 & 0.8 \\ 0.6 & 1 & 0.9 \\ 0.8 & 0.9 & 1 \end{bmatrix}\right]$$

X, P, Q independent standard normal. Results are shown in Figure 6. Again, *MTE* and *IV 1* are the two methods that yield estimates close to the true ones. Here, *MTE* has the larger RMSE. Nevertheless, across all different types of DGPs, *MTE* is the most flexible procedure that allows identification of effects even in situations when classic IV fails.

Figure 6: Results of DGP 2



Notes: Own calculations. This Figure plots the average results of DGP 2 after replicating the estimation respective procedures 200 times. Each panel refers to a specific indirect of direct mediation effect. The green and red horizontal line depict the average true effect for the whole sample and for the Z_1 compliers only (our target parameter), respectively. OLS and IV1 refer to implied mediation effects when $Y = \delta_0 + \delta_D D + \delta_M M + \delta_{DM} DM + \delta_X X + \epsilon$ and $M = \gamma_0 + \gamma_D D + \gamma_X X + u$ are estimated by OLS and two-stage least squares (2SLS), respectively. IV2 reports the implied effect when no interaction term is used in the 2SLS outcome regression. MTE refers to the results of the estimation procedure derived in the paper. Full results including RMSE are reported in Table S1.

DGP 3: Strong heterogeneity between complier groups

While DGP 1 and DGP 2 generate a fairly homogeneous data set—the effects do not vary strongly between complier types—we organize DGP 3 in a way that it produces very heterogenous effects among complier types as well as regarding ITE(1) vs. ITE(0) and DTE(1) vs. DTE(0) in the group of compliers. Additionally, DGP includes a control variable that affects all other observables plus the unobserved complier types. To this end, we define three different DGPs for the three groups (compliers, always takers, and never takers) and then pool the data. In this sense, DGP 3 is more general than DGP 1 and 2, because we not only allow the three groups to have different values of unobservables. In addition, we let the unobservables have an arbitrarily different impact on the outcome across the three groups (holding selection on unobservables constant). Thus, this DGP is in principle more realistic than DGP 1 and 2. The three subsamples are generated as follows:

Complier

$$
\begin{aligned}
Y^{11} &= -5 + 5X + 2U^{11} \\
Y^{10} &= 5 + 5X + 4U^{10} \\
Y^{01} &= -15 + 5X - 2U^{01} \\
Y^{00} &= -5 + 5X - 2U^{00} \\
M^{1} &= \mathbb{1}(3.2 - \gamma_1 + \gamma_2 Z_2 \\
&\qquad -2/9X > U_M^1) \\
M^{0} &= \mathbb{1}(2 - \gamma_1 + \gamma_2 Z_2 \\
&\qquad -2/9X > U_M^0) \\
Z_1 &= \mathbb{1}(P > 0) \\
Z_2 &= \mathbb{1}\{Q > \mathbb{1}\{X > 5\}\} \\
D &= Z_1 \\
X &= \max\{W^C | \\
&\quad 0 \le W^C \le X \le 10\} \\
W^C &= W + \mathbb{1}\{\mathcal{U}(0,1) > 0.85\} \\
W &\sim \mathcal{N}(8,3) \\
P &\sim \mathcal{N}(0,1) \\
Q &\sim \mathcal{N}(0,5.08) \\
\gamma_1 &\sim |\mathcal{N}(0,1)| \\
\gamma_2 &= \gamma_1 + \eta + 0.2 \\
\eta &\sim |\mathcal{N}(0,1)| \\
N &= 20,000
\end{aligned}
$$

Always taker

$$
\begin{aligned}
Y^{11} &= -10 + 5X - 2U^{11} \\
Y^{10} &= -15 + 5X - 4U^{10} \\
M^{1} &= \mathbb{1}(3.2 - \gamma_1 + \gamma_2 Z_2 \\
&\qquad -2/9X > U_M^1) \\
M^{0} &= \mathbb{1}(2 - \gamma_1 + \gamma_2 Z_2 \\
&\qquad -2/9X > U_M^0) \\
Z_1 &= \mathbb{1}(P > 0) \\
Z_2 &= \mathbb{1}\{Q > \mathbb{1}\{X > 5\}\} \\
D &= 1 \\
X &= \max\{W^{AT} | \\
&\quad 0 \le W^{AT} \le X \le 10 \\
W^{AT} &= W - \mathbb{1}\{\mathcal{U}(0,1) > 0.85\} \\
W &\sim \mathcal{N}(8,3) \\
P &\sim \mathcal{N}(0,1) \\
Q &\sim \mathcal{N}(0,5.08) \\
\gamma_1 &\sim |\mathcal{N}(0,1)| \\
\gamma_2 &= \gamma_1 + \eta + 0.2 \\
\eta &\sim |\mathcal{N}(0,1)| \\
N &= 15,000
\end{aligned}
$$

Never taker

$$
\begin{aligned}
Y^{01} &= -15 + 5X - 2U^{01} \\
Y^{00} &= -5 + 5X - 4U^{00} \\
M^{1} &= \mathbb{1}(3.2 - \gamma_1 + \gamma_2 Z_2 \\
&\qquad -2/9X > U_M^1) \\
M^{0} &= \mathbb{1}(2 - \gamma_1 + \gamma_2 Z_2 \\
&\qquad -2/9X > U_M^0) \\
Z_1 &= \mathbb{1}(P > 0) \\
Z_2 &= \mathbb{1}\{Q > \mathbb{1}\{X > 5\}\} \\
D &= 0 \\
X &= \max\{W^{NT} | \\
&\quad 0 \le W^{NT} \le X \le 10\} \\
W^{NT} &= W \\
W &\sim \mathcal{N}(8,3) \\
P &\sim \mathcal{N}(0,1) \\
Q &\sim \mathcal{N}(0,5.08) \\
\gamma_1 &\sim |\mathcal{N}(0,1)| \\
\gamma_2 &= \gamma_1 + \eta + 0.2 \\
\eta &\sim |\mathcal{N}(0,1)| \\
N &= 15,000
\end{aligned}
$$

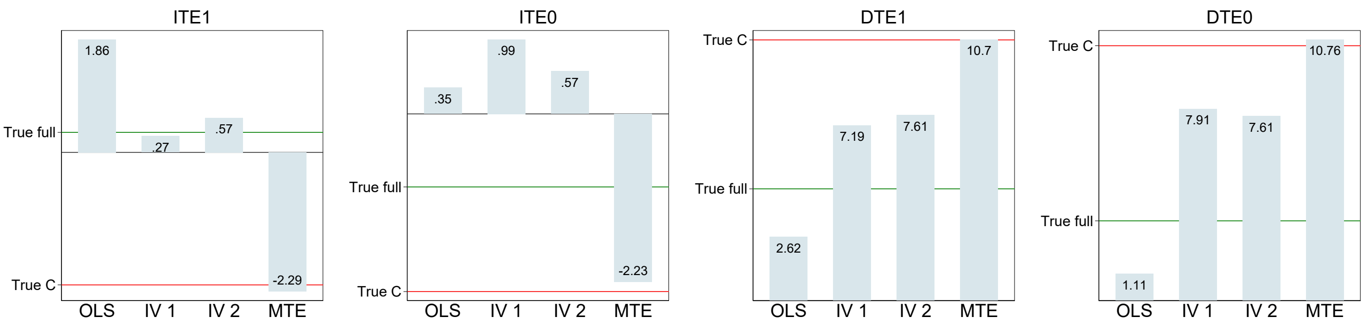
The distribution of the observed confounder X is discrete, ranging between 0 and 10, and depends on the unobserved Z_1 types. In total, X affects all observables. Note that also Z_1 and D implicitly correlate with X, because the complier types have different X values. The unobserved components that affect the potential outcomes of Y and M have a common correlation structure between compliers as well as always and never takers of D. Their distribution reads:

$$\begin{bmatrix} U^{11} \\ U^{10} \\ U^{01} \\ U^{00} \\ U_M^1 \\ U_M^0 \end{bmatrix} \sim \mathcal{N} \left[\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 & 0.8 & 0.8 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 & 0.8 & 0.8 & 0.8 \\ 0.8 & 0.8 & 1 & 0.8 & 0.8 & 0.8 \\ 0.8 & 0.8 & 0.8 & 1 & 0.8 & 0.8 \\ 0.8 & 0.8 & 0.8 & 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 0.8 & 0.8 & 0.8 & 1 \end{bmatrix} \right]$$

As before, $M = M^1 D + M^0(1-D)$ and $Y = DMY^{11} + D(1-M)Y^{10} + (1-D)MY^{01} + (1-D)(1-M)Y^{00}$.

Results are shown in Figure 7. *OLS* and *IV 2* are still far away from the true parameters. Now, however, in this more heterogenous DGP, also classic IV with an interaction of D and M fails. *MTE* now is the only method that produces basically unbiased estimates of direct and indirect effects for compliers in this DGP. This DGP is complex in its type-specific heterogeneity, but it is not unlikely that this is a common feature of real-world data.

Figure 7: Results of DGP 3



Notes: Own calculations. This Figure plots the average results of DGP 3 after replicating the estimation respective procedures 200 times. Each panel refers to a specific indirect of direct mediation effect. The green and red horizontal line depict the average true effect for the whole sample and for the Z_1 compliers only (our target parameter), respectively. OLS and IV1 refer to implied mediation effects when $Y = \delta_0 + \delta_D D + \delta_M M + \delta_{DM} DM + \delta_X X + \epsilon$ and $M = \gamma_0 + \gamma_D D + \gamma_X X + u$ are estimated by OLS and two-stage least squares (2SLS), respectively. IV2 reports the implied effect when no interaction term is used in the 2SLS outcome regression. MTE refers to the results of the estimation procedure derived in the paper. Full results including RMSE are reported in Table S1.

4 Data and institutional set-up

4.1 Sample selection and dependent variable

We use data from the Survey of Health Ageing, and Retirement (SHARE) and the English Longitudinal Study of Ageing (ELSA), two large biennial representative micro data sets providing information on health and other socioeconomic characteristics for individuals aged 50 and older.[9] ELSA started in 2002 with 18,000 individuals while SHARE was initiated in 2004. By now, 8 interview waves of SHARE are available covering information of about 140,000 individuals living in Europe.[10] Both data sets are highly harmonized and can be used for pooled analyses.

For our analysis we use ELSA waves 1–8 and SHARE waves 1, 2, and 4–8 as wave 3 (SHARE-LIFE) treats different aspects and does not contain the variables of interest.[11] We restrict the sample to individuals between 50 and 70 who are at most 8 years above the country-specific early retirement age. In total, we have 80,763 observations from 28,206 individuals living in 8 countries.[12]

Measures of cognitive ability

Cognitive abilities summarize the "ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought" (American Psychological Association, 1995), where the sum of these abilities is referred to as intelligence. SHARE, HRS, and ELSA offer a number of potential measures for cognitive abilities: orientation in time, numeracy, verbal fluency and word recall tests.

[9] This Section heavily draws on text from Schiele and Schmitz (2021) and Schmitz and Westphal (2021).

[10] For comprehensive information on the sampling procedure, questionnaire contents, and fieldwork methodology of HRS, ELSA, and SHARE see Sonnega et al. (2014), Steptoe et al. (2003), and Börsch-Supan and Jürges (2005).

[11] See Börsch-Supan (2019a,b,c,d,e,f,g, 2021); Brugiavini et al. (2019).

[12] These countries are Austria, Germany, Spain, Italy, France, Greece, Czech Republic, and England. Compared to Schneeweis et al. (2014), we additionally include Greece and England. The only country, we do not consider from their study is Denmark, because there was a coinciding reform of the schooling system (the introduction of a comprehensive schooling system) and because there is some disagreement in the literature about the introduction of the reform (i.e. Brunello et al., 2009 report 1971 whereas Arendt, 2005 states the year 1975). From the other possible countries, we do not include Netherlands because of doubts on the enforcement of its relatively small 1950 reform (e.g. in van Kippersluis et al., 2011), Sweden, because school districs could actually decide to implement the reform before 1969 (Lundborg et al., 2014), which makes it hard to detect a sudden and clean jump in years of education without any pre-trends, and Belgium, because there also is some disagreement on the timing of the Belgian reform (the source of Brunello et al., 2016 cannot be verified, Garrouste, 2010 does not report this reform.)

In the *word recall test*, the interviewer reads ten words and the interviewed is asked which of these words they can remember. The number of words they can recall is counted. This word recall test is done twice: directly after the words are read (immediate recall test) and about 5 minutes later (delayed recall test). The total number of words recalled in these two occasions are added up to yield the word recall test score. This score can range between 0 and 20. Word recall is a measure of episodic memory, which is found to react most strongly to aging (Rohwedder and Willis, 2010). It is considered a measure of "fluid intelligence". Broadly speaking, fluid intelligence is the innate cognitive ability while crystallized intelligence is what people learn in their lifetime (using their fluid intelligence).

In the *verbal fluency test* respondents are asked to name as many animals as they can in one minute, where the number of animals they can tell becomes their test score. Here the lower limit is 0, but there is no upper limit (the maximum number in the sample is 100). Verbal fluency is a measure of both fluid and crystallized intelligence as it is both important to know many animals (crystallized knowledge) and to remember them quickly (fluid intelligence). Obviously, both *recall* and *verbal fluency* only capture specific parts of the multidimensional concept "cognitive ability". In our analysis, we follow much of the recent economic literature and employ *recall* as our main variable.[13] It has a mean of 10.41 and a standard deviation of 3.43 in our estimation sample.

4.2 Explanatory variables

Our main explanatory variables are binary measures of education and the current labor force status. We measure education by years of education and define $D = 1$ if the number of completed years of education is at least as large as the compulsory schooling years according to the country-specific rules for the youngest birth cohort in our sample. As an example, if years of compulsory schooling for the birth cohort 1957 in Austria are 9 years, $D = 1$ if individuals have at least 9 years of education (irrespective of the birth cohort). D equals zero if years of education fall below this number. Thus, our binary indicator is a measure of "more" education. Employed as a treatment variable in any IV estimator, this variable enables to condition on those individuals who are affected by compulsory schooling (the compliers). We also report estimates for the total treatment

[13] See, e.g. Rohwedder and Willis (2010) and Celidoni et al. (2017). Mazzonna and Peracchi (2012) and Coe et al. (2012) use recall and a variety of other measures.

effect where years of education are used as an explanatory variable and the results are similar. Yet, for our research design we need a binary treatment.[14]

The upper part of Table 1 shows the distribution of D in our sample and the average number of years of education for observations with $D = 0$ and $D = 1$. Around three fourths of individuals are classified as having more education. This is no surprise as $D = 0$ only for those individuals in countries with changes in compulsory school years who had the mandatory years of schooling according to the old regime. Individuals with more education on average have almost four more years of education.

Table 1: Main explanatory variables

	Observations	Mean years of schooling	Realization of M
Education (D)			
More education (D = 1)	57,044	13.00	
Less education (D = 0)	23,719	9.13	
Labor force status (M)			
Employed/self employed	34,180		0
Unemployed	2,349		0
Retired	34,599		1
Disabled	3,944		1
Not in labor force	5,691		1

Notes: Own calculation based on the pooled selected sample from SHARE and ELSA.

The lower part of Table 1 informs about the labor force status. We treat individuals as being out of the labor force (that is $M = 1$) if they either retired, disabled, or not in the labor force due to other reasons. Individuals who are in the labor force if they either work part- or full time, are self-employed, or choose the response option unemployed in the respective question. The vast majority of individuals is either employed/self-employed or retired. Note that, because we use retirement regulations as an instrument in the subsequent analysis, any effect of M can be equivalently interpreted either generally as dropping out of the labor force or, more specifically, as effects of retirement for the compliers. We vary the definition of M in the robustness checks.

Figure 8 shows cognitive ability (left panel) and labor force participation (right panel) by age and treatment status. Both cognitive abilities and labor force participation strongly decline in age.

[14]Of course, the method could be extended to ordered treatments such as years of education but it is unlikely that existing data sets are large enough to be able to identify parameters in such a model that has a drastically higher demand for data.

Moreover, there are clearly visible correlations with education. At every age, those with more education have by around one unit higher cognitive abilities. This difference seems to increase after the age of 65. In addition, for each age group the share of individuals out of the labor force is smaller among those with more education.

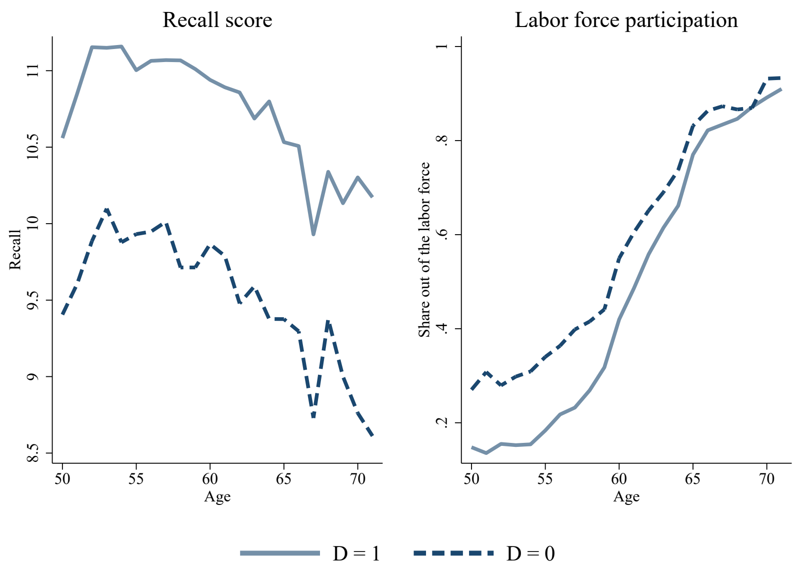


Figure 8: Recall score and labor force participation by age and treatment status
Notes: Own calculation based on the pooled selected sample from SHARE and ELSA. The graph plots unconditional averages by age and treatment status in full years.

In all regressions we use the following control variables: birth year fixed effects, interview wave fixed effects, country fixed effects, country-specific linear age trends, test repetition fixed effects and a gender dummy. Through birth year fixed effects and country fixed effects, we have a difference-in-difference design that enables to compare the arbitrarily set pivotal compulsory schooling cohorts with marginally older ones. These fixed effects are also important for the same reasons as for the retirement effects. As early retirement regulations do not only differ between cohorts but primarily by age, age trends and wave fixed effects are important. These latter controls for age and wave effects are important to differentiate education and retirement effects from the general decline in these age groups. Repetition fixed effects absorb a potential improvement in the recall score that is driven by plain familiarization with the test. In the MTR estimation, we use a slightly different set of variables for the interaction with the propensity score. These include age, male, and test repetitions absorbing level effects of these potentially important variables. Using all variables is infeasible as it would greatly increase the number of estimated parameters leading to problems of overfitting. We show in the robustness checks that our results do not change quantitatively when we omit all interaction terms. Using these interactions, however, the TTE of our MTR estimation recovers the 2SLS-LATE remarkably well.

4.3 Institutional regulations

As is well known, both education and retirement are endogenous when it comes to assessing their impact on cognitive ability. We use two established instruments, compulsory schooling reforms and retirement regulations, in order to identify effects of both variables on the outcome. Compulsory schooling reforms, that is, increases in the mandatory years of education are not free of critique, but are typically considered random from the point of view of the individual, at least conditional on certain control variables.

Table 2: Retirement ages and compulsory schooling

	ERA		Compulsory schooling	
	men	women	change in years	pivotal cohort
Austria	60-65	55-60	8-9	1951
Czech Republic	57-60	54-60	8-9	1934
			9-8	1939
			8-9	1947
England	65-66	60-66	10-11	1957
France	60	60	7-8	1923
			8-10	1953
Germany	63	62-63		
BW			8-9	1953
BY			8-9	1955
HB			8-9	1943
HH			8-9	1934
HE			8-9	1953
NI			8-9	1947
NRW			8-9	1953
RLP			8-9	1953
SL			8-9	1949
SH			8-9	1941
Greece	58-60	55-60	6	1963
Italy	57-58	57-58	5-8	1949
Spain	61	61	6-8	1957

Notes: The table shows for each country and gender the Early Retirement Age (ERA) and for each compulsory schooling reform the change in years of compulsory schooling as well as the first cohort affected by the reform. As ERA depends on e.g. the birth cohort in some countries, we provide the ERA range in our sample for these countries. Information about the compulsory schooling reforms in most countries is taken from Brunello et al. (2016). Additional information about the reforms in Spain, Greece and England is taken from Brunello et al. (2013). Detailed information on retirement rules for each country are in the supplementary materials.

Table 2 reports, for all countries, years of compulsory schooling before and after a regulatory change and the birth cohort that was first affected (that is, the pivotal cohort). In many cases these are increases from 8 to 9 years but there is quite some variation. Some countries witnessed more than one reform and in some countries, specific regions or federal states where subject to different reforms. We take all of this into account. The pivotal cohorts make clear that some reforms were too early or too late to observe both affected and not affected individuals within the same country. This is discussed in more detail below. We define the instrument by pooling the information of the compulsory schooling reforms and set $CS = 1$ for the pivotal cohorts and those born later and $CS = 0$ else.

Table 2 also reports early retirement ages (ERA), which are used as instruments for being retired. The early retirement age is the age individuals are allowed to retire for the first time (as long as their are not disabled). Early retirement goes along with a penalty on the retirement benefits, which usually is gradually decreased until the official retirement age, (ORA) is reached. The institutional rules used to calculate early retirement ages are reported in the supplementary materials. Early retirement ages mainly vary by country, gender and over time. However, in part they also vary by the individual work history and year of birth. There is a slight tendency to increase the early retirement age over time within the countries as a reaction of social policies to the challenges brought along by the demographic change, which we exploit as one source of exogenous variation.

Retirement regulations are a common instrument in the literature, see, e.g. Celidoni et al. (2017), Mazzonna and Peracchi (2012), Mazzonna and Peracchi (2017). However, in contrast to the previous literature, we only use the *early* but not the *official* retirement age as an instrument and define $ERA = 1$ if the early retirement age is reached and $ERA = 0$ if not. Please refer to Schmitz and Westphal (2021) who show that—at least in these data—the first-stage effect of reaching the ORA is not strong enough to use it as an instrument, once age is properly controlled for. It turns out that ERA seems to be the more important incentive to retire than ORA. We think about the ERA as the onset of a dynamic incentive structure that also comprises the ORA. We report results of additionally including ORA as an instrument for M in the robustness checks.

Table 3 shows the interplay of both instruments ERA and CS. For identification we need individuals born before and after the pivotal cohorts and—within both groups—individuals above and below early retirement age. Take, as one example, Austria, the first line of Table 2. We have data for the years 2004–2020 (the ELSA data spans until 2018). In 2016, the pivotal cohort of 1951 turned 65, the (then) ERA for men. Thus, (only) in the most recent wave of data we observe men

Table 3: When can the pivotal cohorts retire?

Country	Pivotal Cohort	Year surveyed in SHARE/ELSA: Men 04	06	08	10	12	14	16	18	20	Women 04	06	08	10	12	14	16	18	20
Austria	1951	x	x	x	x	x	x	x	o	o	x	x	x	o	o	o	o	o	o
Czech Republic	1934	x	x	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
	1939	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
	1947	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
England	1933	o	o	o	o	o	o	o	o	–	o	o	o	o	o	o	o	o	–
	1957	x	x	x	x	x	x	x	x	–	x	x	x	x	x	x	x	o	–
France	1923	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
	1953	x	x	x	x	x	o	o	o	o	x	x	x	x	x	o	o	o	o
Germany																			
BW	1953	x	x	x	x	x	x	o	o	o	x	x	x	x	x	x	o	o	o
BY	1955	x	x	x	x	x	x	x	o	o	x	x	x	x	x	x	x	o	o
HB	1943	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
HH	1934	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
HE	1953	x	x	x	x	x	x	o	o	o	x	x	x	x	x	x	o	o	o
NI	1947	x	x	x	o	o	o	o	o	o	x	x	x	o	o	o	o	o	o
NRW	1953	x	x	x	x	x	x	o	o	o	x	x	x	x	x	x	o	o	o
RLP	1953	x	x	x	x	x	x	o	o	o	x	x	x	x	x	x	o	o	o
SL	1949	x	x	x	x	o	o	o	o	o	x	x	x	x	o	o	o	o	o
SH	1941	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
Greece	1963	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Italy	1949	x	o	o	o	o	o	o	o	o	x	o	o	o	o	o	o	o	o
Spain	1957	x	x	x	x	x	x	x	o	o	x	x	x	x	x	x	x	o	o

x stands for: not yet old enough to be eiligible to retire
o stands for: old enough to be eiligible to retire
– stands for: no data

Notes: Own illustration based on information in Table 2.

from Austria with $CS = 1$ and $ERA = 1$. We do, however, also observe all three other combinations of CS and ERA. Since the ERA is 60 in 2016 for women, we have more observations with $CS = 1$ and $ERA = 1$ here and, again, also observations with all three other combinations. In the Czech Republic, as another example, compulsory schooling reforms took place so early that we do not observe individuals with $CS = 1$ and $ERA = 0$. Thus, we, in parts, need to rely on cross-country variation in institutional settings.

The number of observations for each country and by the treatment D (education), the mediator M (being in the labor force) and both instruments are reported in Table A1. Note that we have three countries, UK, Greece, and Spain, where the respective compulsory schooling reform was too late for individuals in our sample such that no one older than 55 was affected. Those three countries serve as a control group. Having a significant fraction of observations where the instrument is never switched on may help to prevent important drawbacks of difference-in-differences and event-study settings (Sun and Abraham, 2021; De Chaisemartin and D'HaultfŒuille, 2018; De Chaisemartin and d'Haultfoeuille, 2020).[15] We test whether our results are affected by these problems in the robustness checks.

5 Results

5.1 Total Treatment Effect

We start with the (local average) total treatment effect. That is, the effect of education in adolescence on cognitive abilities later in life without studying any pathways such as labor force participation. We first present some event-study evidence on the first stage and the reduced form. Both analyses estimate cohort-specific effects relative to the pivotal cohort and its relationship with schooling. The first stage identifies the complier share for the corresponding event years, while the reduced form gives the respective effects on cognitive abilities. Both analyses rely on the following regression, where the event time r is the normalized birth year ($r =$ birth year – pivotal cohort)

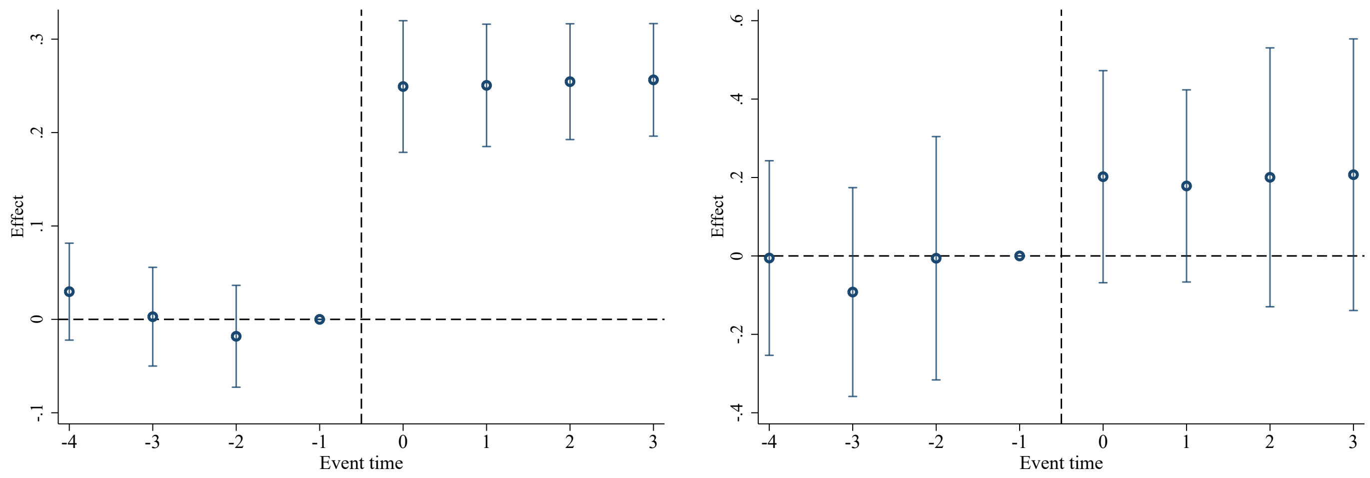
$$Y_{it} = \gamma_{pre} \mathbb{1}[r < -4] + \sum_{\substack{j=-4 \\ j \neq -1}}^{\infty} \gamma_j \mathbb{1}[r = j] + \boldsymbol{X}_{it}' \boldsymbol{\beta} + \varepsilon_{it},$$

where there vector $\boldsymbol{X}_{it}$ consists of the controls mentioned in the previous section. For the first stage, we use the education indicator D_{it} as the dependent variable, whereas the reduced form uses cognitive abilities Y_{it}. The coefficients of interest are γ_1, γ_2, and γ_3, which measure the effects of compulsory schooling for the different cohorts relative to $r = -1$ (our reference category).

[15] These drawbacks could arise in linear models because of an overprediction of the probability to take the treatment (being affected by compulsory schooling) by combining the manifold fixed effects that are often necessary for identification (in our application especially cohort and country fixed effects). Observations at the end of the observation period (the youngest cohorts) in groups that are treated early (countries that introduced compulsory schooling first) are particularly likely to have a predicted first stage value that exceeds one (through the combination of the corresponding fixed effects). De Chaisemartin and d'Haultfoeuille (2020) showed that this overprediction leads to a negative weight for the treatment effect of the corresponding units. Including countries without a compulsory schooling reform in the considered period alleviate these problems.

Additionally, we are interested in the pre-event coefficients γ_{-4}, γ_{-3}, and γ_{-2} as these detect any deviation from a common trend in the outcome and treatment prior to the introduction of the reform. We plot these coefficients in Figure 9. We do not plot coefficients for higher event times to ensure that a homogeneous sample of countries contributes to the effects. Nonetheless, we include saturated event-time effects for all periods larger than 3 (up to $r = 22$) to avoid a contamination from other periods. The same holds for event times less than -4. For these, we include a joint indicator.

Figure 9: Compulsory schooling: first stage and reduced form on cognitive abilities



(a) First Stage: Compulsory schooling

(b) Reduced form: Compulsory schooling

Notes: Event time is birth cohort minus pivotal cohort. The vertical lines around the dots indicate 95 percent confident intervals. Standard errors clustered on birth year-country level.

In the left panel of Figure 9, we present the first stage, which is estimated precisely. There are no pre-trends and a clearly visible and persistent upward shift in education when a compulsory schooling reform took effect. The jump of the effects between $r = -1$ and $r = 0$ means that, on average, 24 percent of all individuals had to extend their schooling because of the compulsory schooling reforms.[16] This is the group of compliers in the IV terminology. Although being a rather small group of individuals, it is a potentially very interesting one: compliers in our case are those who only take the minimum necessary level of education. Thus, we will identify effects for individuals with a low preference for education.

Figure 9b shows the effects on cognitive abilities. In the four years prior to the reform, there are no differential effects visible, suggesting our estimates later do not capture some specific cohort trends. Starting with the first affected compulsory-schooling cohort, we see an elevated level

[16]The exact number of 24 percent follows from Table 4 below.

of cognitive skills, which stays around 0.2 (amount of words recalled more) for the subsequent periods. Although, when taken on its own, neither post compulsory-schooling effect is significant at the 95% significance level, a joint estimation reveals a precise reduced-form effect, which is presented below. In total, the event-study evidence suggest that the compulsory schooling reforms sharply increased years of schooling for the affected cohorts without any detectable pre-trends. Thus, these changes most likely explain similar old-age cognitive ability trajectories for cohorts around the introduction of the reforms.

Table 4 reports regression results of both OLS and 2SLS (including first stage) of cognitive abilities on education and controls. OLS in column (1) repeats the finding from Figure 8 that already showed a higher level of cognitive abilities for those with more education. Those with more education score, on average, almost 1.4 points higher in the cognitive abilities test. The first stage regression of D on the instrument Z_1 in column (2) aggregates the numbers already seen in Figure 9a to 0.24, i.e. 24 percent of all individuals are compliers because they need to adjust their otherwise preferred amount of schooling due to the compulsory schooling reform. We can contrast this finding with the reduced-form result in column (4). Compulsory schooling raises number of recalled words in the overall population by 0.2. This effect is precisely estimated and significant at the 5 percent level. Finally, the effect of more education on cognitive abilities in column (3) is around 0.8. The compliers to the compulsory schooling reforms recall 0.8 more words due to more education. This is a considerable amount. To interpret this effect size, we could, for instance, relate it to the level or the trend plotted in Figure 8. The individuals with $D = 0$ are a more appropriate comparison, as these are all potential compliers with the same level of schooling that also the treated compliers would have preferred. Compulsory schooling could raise the level of cognitive abilities around age 55-60 by about 8 percent, or by 24 percent of a standard deviation in the full sample. Another way of interpretation is the comparison to the general age-related decline, which is about one word in 15 years, see Figure 8. Thus, compulsory schooling could roll back the age-related decline in recall at age 70 by approximately 12 years. In Table A2 in the Appendix, we add regression results with a different treatment variable, namely years of education. The results are comparable.

Table 4: Regression results: total treatment effect

	OLS (1)	First stage (2)	2SLS (3)	Reduced Form (4)
More education (D)	1.423^{***}		0.811^{***}	
	(0.052)		(0.306)	
Post CS-reform (Z_1)		0.244^{***}		0.198^{**}
		(0.020)		(0.081)
Control variables	yes	yes	yes	yes

Number of observations in each regression: 80,763. Additional control variables are birth year fixed effects, interview wave fixed effects, country fixed effects, country-specific linear age trends, test repetition fixed effects and male. Standard errors in parentheses clustered on birth year-country level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

5.2 Effect of education on retirement

We now estimate the effect of education on retirement. Such an effect is a prerequisite for an indirect effect of education on cognitive abilities through the channel of labor force participation. As Table 5 shows, there is indeed a considerable effect. The compliers in our sample are by 17.7 percentage points less likely to be retired due to compulsory schooling. This effect is even larger than the OLS difference, demonstrating that particularly individuals at the lowest educational margins adjust their labor supply in response to more education.

Table 5: Regression results: Effect of D on M

	OLS (1)	2SLS (2)
More education (D)	-0.082^{***}	-0.177^{***}
	(0.008)	(0.050)
Control variables	yes	yes

Number of observations in each regression: 80,763. Additional control variables are birth year fixed effects, interview wave fixed effects, country fixed effects, country-specific linear age trends, test repetition fixed effects and male. Standard errors in parentheses clustered on birth year-country level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Moreover, the effect of retirement on cognition is well-documented, see, e.g. Schmitz and Westphal (2021). Their study also shows that there are no pre-trends neither in cognitive abilities nor in the probability to be in the labor force prior to retirement eligibility.

5.3 Evidence on the identifying assumptions and other threats to the validity of the estimates

Before we can decompose the local average treatment effect of education on cognitive decline, we check the validity of the additional assumptions for the mediation analysis and other general threats. First, we gauge the conditional independence of the instruments (Assumption 4). One testable implication of this assumption is that information on one instrument does not help in predicting the other. Table A3 suggests that this is the case: individuals who are affected by compulsory schooling cannot retire at different ages conditional on control variables compared to individuals who are not. Second, we can show that we have full support in the propensity score of M (Assumption 5). Together with the assumption that the observed and unobserved heterogeneity in the outcome Y are additively separable (Assumption 2), this enables estimating marginal treatment effects over the whole unit interval. Figure A2 depicts the resulting distribution of the propensity score by labor force status (complete estimation results are presented in Table A4) . In total, the common support ranges from 0.026 to 0.991—nearly of the whole unit interval. Hence, the resulting TTE from our unconstrained MTE approach should recover the TTE estimated by 2SLS in Table 6.

Third, we want to scrutinize whether our settings suffers from problems in two-way fixed effects setting, as demonstrated, for instance, in Sun and Abraham (2021). The authors demonstrate that conventional event-study estimates could get contaminated with effects from other relative time periods. Figure A3 plots the results of their proposed estimator that prevents such a contamination. It shows that the first-stage and the reduced-form effects are both somewhat larger when accounting for this contamination. The difference is small, however, leaving the general conclusion unchanged. Moreover, the difference reduces for the IV estimates (which is the ratio of reduced form and first stage).

5.4 Estimated potential outcomes and treatment effects

Now we can turn to our main results—a formal mediation analysis of the effect of education on cognitive abilities that controls labor force participation as a potential mediator. To this end, we present the total treatment effect (TTE) and the two respective indirect and direct treatment effects in Table 6. Subsequently, we discuss the underlying components that constitute these parameters (the MTR curves) in more detail. Turning to Table 6, the the first line (labeled as MTE) are the main results of this paper. We contrast these estimates with those of three other estimators. First, we employ a slightly adjusted estimator of our MTE approach, where we estimate the ITEs as before, but use 2SLS to estimate the TTE and then derive the DTEs using Eq. (1). By exploiting this restriction, we may gain efficiency for estimating the TTE and the DTEs. Second, we employ a conventional IV approach (referred to as IV 1 in Section 3). Lastly, we ignore endogeneity of D and M and estimate these parameters by OLS for completeness (also described in Section 3). Concerning the estimated results, the TTE amounts to 0.804–0.864 depending on the estimator (we ignore OLS because it is inconsistent). The difference in the TTE between the MTE and 2SLS estimator is remarkably small and negligible. This is reassuring that our estimator does well in estimating the local average treatment effect. A relevant difference is the larger standard errors, which is due to a more inefficient estimation with many estimated parameters more than necessary for the LATE in three different regressions. Nonetheless, the MTE-TTE still is significant on the 10 percent level.

Table 6: Main results—total, direct, and indirect treatment effects

	Total treatment effect	Effect decomposition			
		Indirect TEs		Direct TEs	
	$TTE = LATE$	$ITE(1)$	$ITE(0)$	$DTE(1)$	$DTE(0)$
MTE	0.864*	0.293*	0.043	0.822*	0.571
	(0.505)	(0.153)	(0.052)	(0.494)	(0.488)
MTE (derived)	0.811**	0.293*	0.043	0.768**	0.518
	(0.380)	(0.153)	(0.052)	(0.377)	(0.395)
2SLS	0.804**	0.184	−0.209	1.013**	0.620*
	(0.379)	(0.140)	(0.164)	(0.485)	(0.354)
OLS	1.418***	0.039***	0.050***	1.368***	1.380***
	(0.052)	(0.005)	(0.009)	(0.054)	(0.054)

Number of observations: 80,763. Control variables are birth year fixed effects, interview wave fixed effects, country fixed effects, country-specific linear age trends, test repetition fixed effects and male. Bandwidth = 0.25. Bootstrap standard errors (200 replications) in parentheses clustered on birth year-country level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Turning to the effect decomposition, recall that by construction $ITE(1)$ and $DTE(0)$ as well as $ITE(0)$ and $DTE(1)$ each form a pair that each add up to the total treatment effect. $ITE(1)$ amounts to 0.293 for the MTE estimator (by definition, MTE and MTE (derived) are equivalent for the ITEs). Thus, $ITE(1)$ amounts to more than one third of the total effect and is significantly estimated at the 10 percent level. Estimating this effect by 2SLS yields a much smaller (and probably biased) estimate. In our application, $ITE(1)$ measures the contribution of a changed labor force participation caused by compulsory schooling for the more educated (treated) compliers. Education made these individuals work to older ages in their jobs (which are probably of a different quality than for less-educated compliers) before retiring. This quantitatively meaningful parameter documents that the TTE is rather not constant over the life course but may arise through individual decisions after education is finished—in particular, decisions on the jobs and the retirement timing of these individuals.

The remaining part for the TTE is $DTE(0)$, i.e. the effect of increasing education due to compulsory schooling when labor force participation is fixed at M^0—the hypothetical labor force participation for compliers without the compulsory schooling reform. This effect amounts to 0.571 without a restriction and 0.518 when we use the TTE and ITE(1) to derive the DTE. Both effects are not statistically significant at any conventional level. However, the magnitude of the effect is not negligible. In any case, given that the DTE includes the direct effect of education plus all other mediating forces that are unrelated to labor force participation, it again outlines the relevance of labor force participation for the education effect.

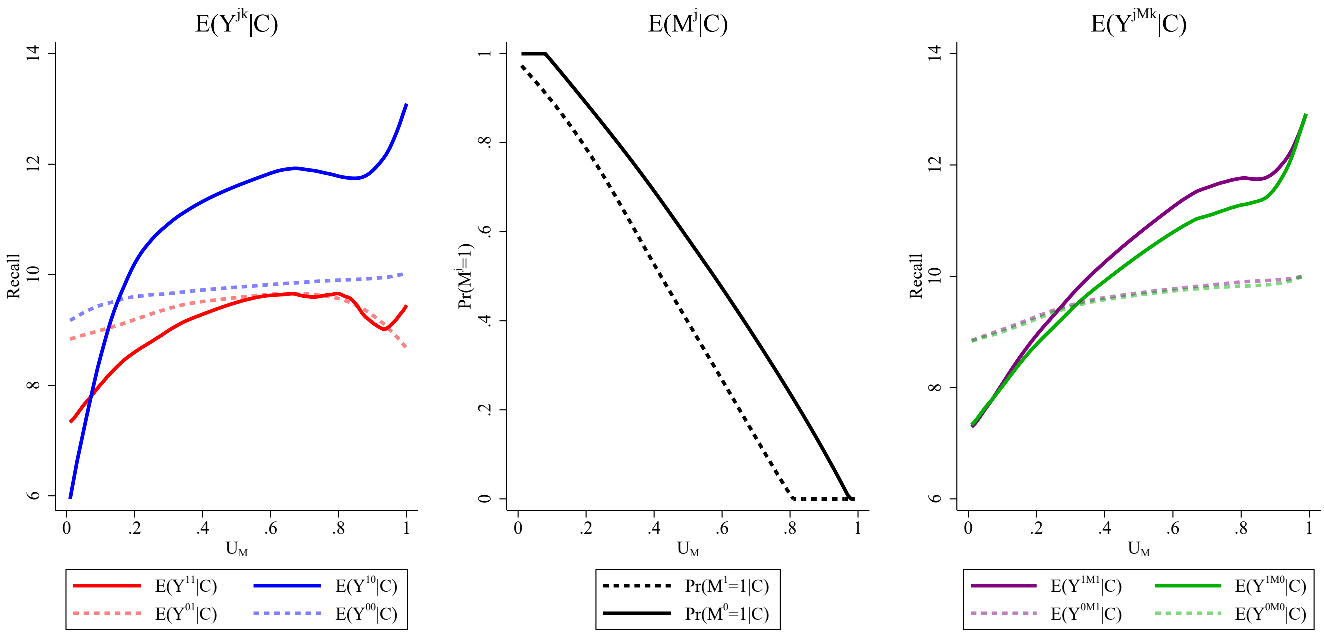
Turning to the other pair, $ITE(0)$ and $DTE(1)$, a slightly different picture emerges. We see that the indirect effect of labor force participation is much smaller (a contribution of 0.043 words to the TTE) and insignificant. Only slightly more than five percent of the TTE runs through an effect of retirement on cognitive abilities for the compliers with less education. The difference between $ITE(0)$ and $ITE(1)$ is that the former is the effect of retirement on cognitive abilities for individuals with lower education (as opposed to the more-educated compliers for which the $ITE(1)$ applies). This demonstrates an important complementary effect between education and labor force participation. Only more educated compliers gain through labor force participation, likely because their job environment is more stimulating. In contrast, less educated compliers do not gain through working to older ages. The lower $ITE(0)$ mechanically forces $DTE(1)$ also to be lower than $DTE(0)$. To provide an explanation for this result, consider the definition of $DTE(1) = E(Y^{1M^1} - Y^{0M^1})$. It is the causal effect of education on cognitive abilities if the individual retirement behavior was like the one of the more educated individuals. Because more educated individuals retire later, they are longer in arguably more stimulating environments where they can sustain a higher level of cognitive abilities more easily. Hence, this also shows the complementarity between education and labor force participation.

Now, we want unravel our results by showing them along the margin of indifference for retirement (along the index U_M) because this is key for our approach to work. To do this, we plot the four MTR functions—the average level of cognitive abilities by the education dummy D and the labor force dummy M along U_M for compulsory schooling compliers—in the right panel of Figure 10. For the sake of visibility, we ignore the standard errors of these lines. It shows that relative to the unconditional mean of the recall score in our sample ($\overline{Y_{it}} = 10.41$), almost all MTRs for all values of U_M are lower. Only some compulsory schooling compliers with more education ($D = 1$) and a low preference for retiring (high U_M) who also do not retire ($M = 0$) are still in the labor force have higher cognitive scores than the overall average. This is unsurprising, as only compulsory schooling compliers at the lowest educational margins contribute to the plotted MTRs. Moreover, we can see that almost across the whole unit interval, $E(Y^{10}|U_M = p, Z_D \text{ compliers})$ is dominating the other MTRs. Among this group, only individuals with the highest retirement preference $U_M < .2$ have actually one of the lowest average recall scores. The second highest ability curve is for individuals still in the labor force who have years of schooling lower than the new compulsory standard ($E(Y^{01}|U_M = p, Z_D \text{ compliers})$). Given that both curves for labor force participating individuals are the highest emphasizes the role of work-related cognitive stimulation in maintaining the cognitive abilities at older ages. The remaining two curves for individuals who are out of the labor force are lower. Somewhat surprisingly perhaps, education does not seem to have a consistently positive effect for individuals who are retired (the difference between the red MTRs). This finding may suggest that employment and schooling are complementary for maintaining the cognitive abilities emphasizing again the potential role of labor force participation as a mediating channel. We now focus on this indirect effect more formally.

The middle panel of Figure 10 informs about the retirement probabilities by education indicator D for compulsory schooling compliers ($Pr(M^1 = 1|C, U_M)$ and $Pr(M^0 = 1|C, U_M)$). These quantities both decline mechanically along U_M. It is also clearly visible that at all U_M values, the probability to be retired is lower for the more educated compliers. This means that the indirect treatment effect is composed of individuals with all possible retirement preferences.

The last panel presents the mediated outcomes Y^{jM^l}, which are aggregated curves from the first two panels by weighting the four potential outcomes by the potential retirement probabilities for more and less educated compliers.[17] Average differences between any two of those lines give a certain mediation effect. For instance, the differences between the solid (dashed) purple and green lines yield the $ITE(1)$ ($ITE(0)$). The difference between the purple (green) solid and dashed line gives the $DTE(1)$ ($DTE(0)$). This figure, in particular the left panel, alludes to the primary cause of differences in the mediation effects. It is the high level of cognitive abilities of more educated compliers who are still in the labor force who are causing the the TTE in general, but also the differences between the ITEs, in particular. This finding applies to individuals with almost all retirement preferences—except for the very few with the highest preference. This emphasizes once again the complementarity between education and labor force participation.

Figure 10: Estimated marginal treatment response functions for compulsory-schooling compliers



Notes: Number of observations: 80,763. Control variables are birth year fixed effects, interview wave fixed effects, country fixed effects, country-specific linear age trends, test repetition fixed effects and male.

Before we conclude, Table 7 documents the robustness of our estimates with respect to some meaningful changes in the estimation procedure (i.e., the size of the bandwidth), the sample composition, and the the retirement definition. Concerning the bandwidth in the nonparametric MTR estimation, the effects barely change quantitatively. It does not make a difference if we additionally instrument labor force participation by an indicator for the official retirement age (ORA), as is

[17]The treatment-specific quantities are related as follows: $Y^{jM^j} = Y^{j1}M^j + Y^{j0}(1 - M^j)$ for the observed and complier-specific outcome and $Y^{jM^i} = Y^{j1}M^i + Y^{j0}(1 - M^i)$, where $j \neq 0$, for the counterfactual complier-specific outcome if the labor force participation would be manipulated to the other treatment state.

sometimes done in the literature on the effects of retirement. Dropping the unemployed, disabled or homemaker increases all mediation parameters, but qualitatively the results are the same (except that now, the ITE(0) is also meaningful in magnitude). Changing the retirement definition to include the unemployed has no visible implications for the magnitude of the effects. Stratifying the effects by gender, however, outlines important heterogeneities in the effects. Whereas males have a smaller TTE compared to females, their ITE(1) is considerably larger. This may be due to the fact that males typically have been more career-oriented in the past, which includes working until statutory retirement ages. When retiring, more will change for them and the drop in the level of stimulation of their cognitive abilities may thus be larger. Education, in turn, generally may change more for females apart of their working environment, including working at all (even if small hours), finding a different partner, etc. Finally, we scrutinize the results with respect to the compulsory schooling reforms. Keeping only countries in which the compulsory schooling increase was one year, may create a more homogeneous sample of countries. This attenuates the TTE somewhat and reduces the precision of all parameters, but the magnitude of all effects (except ITE(0), as before) remains meaningful. In total, we confirm the robustness of our results in some important dimensions suggesting that our results are not driven by some arbitrary choices in the sample selection, treatment definition or estimation procedure.

6 Conclusion

We study the interaction of education in adolescence and labor-force participation around retirement age and its effect on cognitive abilities of individuals in Europe aged 55-70. Our main goal is to separate the total effect of education on older-age cognitive abilities into a direct effect and indirect effect through labor-force participation. By this, we aim at putting the results found in the literature so far in a more consistent perspective—as we believe that the age gradient in the effects (see introduction) may be caused by downstream differences in the cognitive environment (such as differences in labor force participation around retirement ages) long after education is finished. To this end, we conduct a causal mediation analysis. Since both education and retirement are subject to individual choice (and, thus, endogenous), we exploit exogenous variation from compulsory schooling reforms and early retirement regulations for identification. We demonstrate how the marginal treatment effects framework can be used to conduct such a causal mediation analysis

Table 7: Robustness and other specifications

	Total treatment effect	Effect decomposition				
		Indirect TEs		Direct TEs		
	$TTE = LATE$	$ITE(1)$	$ITE(0)$	$DTE(1)$	$DTE(0)$	N
Bandwidth = 0.15	0.857*	0.243	0.058	0.799*	0.614	80,763
	(0.466)	(0.179)	(0.055)	(0.455)	(0.459)	
Bandwidth = 0.2	0.923*	0.284*	0.049	0.875*	0.639	80,763
	(0.480)	(0.161)	(0.054)	(0.468)	(0.467)	
Bandwidth = 0.3	0.811	0.281*	0.038	0.773	0.530	80,763
	(0.501)	(0.147)	(0.050)	(0.489)	(0.478)	
Bandwidth = 0.35	0.809*	0.269*	0.035	0.775	0.540	80,763
	(0.485)	(0.144)	(0.050)	(0.474)	(0.462)	
With ERA and ORA as instruments for M	0.936*	0.241	0.018	0.918*	0.695	80,763
	(0.507)	(0.153)	(0.044)	(0.501)	(0.501)	
Without unemployed, disabled, homemakers	1.132**	0.398*	0.153***	0.980**	0.734*	68,779
	(0.489)	(0.205)	(0.048)	(0.475)	(0.417)	
Unemployed = retired	0.967*	0.290	0.009	0.958*	0.677	80,763
	(0.544)	(0.179)	(0.062)	(0.543)	(0.533)	
Male	0.582	0.552**	0.099	0.483	0.0294	43,397
	(0.659)	(0.261)	(0.0754)	(0.631)	(0.612)	
Female	1.148	0.365	0.0274	1.121	0.783	37,366
	(0.827)	(0.309)	(0.0677)	(0.811)	(0.665)	
One year increase in CS	0.729	0.365	0.003	0.726	0.364	59,635
	(0.816)	(0.272)	(0.043)	(0.817)	(0.818)	

Control variables are birth year fixed effects, interview wave fixed effects, country fixed effects, country-specific linear age trends, test repetition fixed effects and male. ERA (ORA) refer to indicators of being above the early (official) retirement age in the respective country. Bootstrap standard errors (200 replications) in parentheses clustered on birth year-country level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

that accommodate heterogeneous treatment effects and non-compliance in educational as well as labor supply decisions.

We pool data from SHARE and ELSA on 80,000 observations from several countries in Europe across the years 2002-2020. The data include experimentally collected measures of cognitive abilities (word recall test and verbal fluency test). In a first step, we are able to replicate the effect of education on cognitive abilities as found in the literature. When we split up this effect into a direct effect of education and an indirect effect through labor-force participation, we can show that retirement may be crucial for the onset of a cognitive decline. Retirement behavior for more-educated compulsory schooling compliers may explain more than one third of the total effect. Moreover, we find evidence that schooling and labor supply are complementary, i.e. more education and labor-force participation together seem to protect most against a cognitive decline. In total, these results may explain the heterogeneous effect patterns that other studies found and which could be deemed as inconsistent when not accounting for labor supply as a mediator.

Of course, later-life labor-force participation is only one of a multitude of potential mediators of the effect of education on cognitive decline and not necessarily the most important one. Occupational choice and health behavior probably are two others that directly come to mind. Moreover, middle-life education, middle-life labor-force participation, family status and other forms of cognitive stimulation are likely to play a role. Thus, this analysis, even if claimed to be a causal mediation analysis is only able to inform about a small detail of the bigger picture "cognitive decline". What is called "direct effect" is, as obviously in all mediation analyses only a compound one of the actual direct effect and other not measured indirect effects. Future work might simultaneously take into account more indirect paths of education on cognitive abilities—with the increased demand for data and exogenous variation that comes along with this.

Nonetheless, important policy implications may arise. Policy could act today to still reap effects of past education reforms by enabling individuals to maintain a more cognitive stimulating environment until older ages. More liberal retirement policies and flexible work arrangements appear as important instruments for this that would come at almost no cost. Our results also demonstrate important side effects of education, which are not detectable shortly after education is completed, but emerge over the life course. This life-course perspective needs to be taken into account also when assessing the non-monetary benefits of education.

Acknowledgments

SHARE: This paper uses data from SHARE Waves 1, 2, 3, 4, 5, 6, 7 and 8 (DOIs: 10.6103/SHARE.w1.710, 10.6103/SHARE.w2.710, 10.6103/SHARE.w3.710, 10.6103/SHARE.w4.710, 10.6103/SHARE.w5.710, 10.6103/SHARE.w6.710, 10.6103/SHARE.w7.711, 10.6103/SHARE.w8.100, 10.6103/SHARE.w8ca.100), see Börsch-Supan et al. (2013) for methodological details. The SHARE data collection has been funded by the European Commission, DG RTD through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7

(SHARE-PREP: GA N°211909, SHARE-LEAP: GA N°227822, SHARE M4: GA N°261982, DASISH: GA N°283646) and Horizon 2020 (SHARE-DEV3: GA N°676536, SHARE-COHESION: GA N°870628, SERISS: GA N°654221, SSHOC: GA N°823782) and by DG Employment, Social Affairs & Inclusion through VS 2015/0195, VS 2016/0135, VS 2018/0285, VS 2019/0332, and VS 2020/0313. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C, RAG052527A) and from various national funding sources is gratefully acknowledged (see www.share-project.org).

ELSA: This analysis uses data or information from the Harmonized ELSA dataset and Codebook, Version E as of April 2017 developed by the Gateway to Global Aging Data. The development of the Harmonized ELSA was funded by the National Institute on Aging (R01 AG030153, RC2 AG036619, 1R03AG043052). For more information, please refer to www.g2aging.org.

References

American Psychological Association (1995). Intelligence: Knowns and Unknowns, Report of a task force convened by the American Psychological Association. *Science Directorate, Washington DC [LGH]*.

Arendt, J. N. (2005). Does education cause better health? a panel data analysis using school reforms for identification. *Economics of Education Review*, 24(2):149–160.

Atalay, K., Barrett, G. F., and Staneva, A. (2019). The effect of retirement on elderly cognitive functioning. *Journal of Health Economics*, 66:37 – 53.

Banks, J. and 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.

Banks, J., O'Dea, C., and Oldfield, Z. (2010). Cognitive Function, Numeracy and Retirement Saving Trajectories. *Economic Journal*, 120(548):381–410.

Banks, J. and Oldfield, Z. (2007). Understanding Pensions: Cognitive Function, Numerical Ability and Retirement Saving. *Fiscal Studies*, 28(2):143–170.

Bonsang, E., Adam, S., and Perelman, S. (2012). Does retirement affect cognitive functioning? *Journal of Health Economics*, 31(3):490–501.

Börsch-Supan, A. (2019a). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 1. Release version: 7.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w1.700. Technical report.

Börsch-Supan, A. (2019b). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 2. Release version: 7.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w2.700. Technical report.

Börsch-Supan, A. (2019c). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 3 - SHARELIFE. Release version: 7.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w3.700. Technical report.

Börsch-Supan, A. (2019d). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 4. Release version: 7.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w4.700. Technical report.

Börsch-Supan, A. (2019e). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 5. Release version: 7.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w5.700. Technical report.

Börsch-Supan, A. (2019f). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 6. Release version: 7.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w6.700. Technical report.

Börsch-Supan, A. (2019g). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 7. Release version: 7.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w7.700. Technical report.

Börsch-Supan, A. (2021). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 8. Release version: 1.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w8.100. Technical report.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., and Zuber, Sabrina, o. b. o. t. S. C. C. T. (2013). Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*, 42(4):992–1001.

Börsch-Supan, A. and Jürges, H. (2005). The Survey of Health, Aging and Retirement in Europe – Methodology. Technical report, Mannheim: Mannheim Research Institute for the Economics of Aging.

Brinch, C. N. and Galloway, T. A. (2012). Schooling in adolescence raises iq scores. *Proceedings of the National Academy of Sciences*, 109(2):425–430.

Brinch, C. N., Mogstad, M., and Wiswall, M. (2017). Beyond late with a discrete instrument. *Journal of Political Economy*, 125(4):985–1039.

Brugiavini, A., Orso, C. E., Genie, M. G., Naci, R., and Pasini, G. (2019). Combining the retrospective interviews of wave 3 and wave 7: the third release of the share job episodes panel. share working paper series: 36-2019. munich: Mea, max planck institute for social law and social policy. Technical report.

Brunello, G., Fabbri, D., and Fort, M. (2013). The Causal Effect of Education on Body Mass: Evidence from Europe. *Journal of Labor Economics*, 31(1):195–223.

Brunello, G., Fort, M., and Weber, G. (2009). Changes in compulsory schooling, education and the distribution of wages in europe. *The Economic Journal*, 119(536):516–539.

Brunello, G., Weber, G., and Weiss, C. T. (2016). Books are Forever: Early Life Conditions, Education and Lifetime Earnings in Europe. *The Economic Journal*, 127(600):271–296.

Carlsson, M., Dahl, G. B., Öckert, B., and Rooth, D.-O. (2015). The effect of schooling on cognitive skills. *Review of Economics and Statistics*, 97(3):533–547.

Carneiro, P., Heckman, J. J., and Vytlacil, E. J. (2011). Estimating Marginal Returns to Education. *American Economic Review*, 101(6):2754–81.

Carneiro, P. and Lee, S. (2009). Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality. *Journal of Econometrics*, 149(2):191–208.

Celidoni, M., Bianco, C. D., and Weber, G. (2017). Retirement and cognitive decline. a longitudinal analysis using share data. *Journal of Health Economics*, 56:113 – 125.

Chandra, A., Coile, C., and Mommaerts, C. (2022). What Can Economics Say About Alzheimer's Disease? *Journal of Economic Literature*, forthcoming.

Chen, S. H., Chen, Y.-C., and Liu, J.-T. (2019). The Impact of Family Composition on Educational Achievement. *Journal of Human Resources*, 54(1):122–170.

Chen, Y.-T., Hsu, Y.-C., and Wang, H.-J. (2020). A stochastic frontier model with endogenous treatment status and mediator. *Journal of Business & Economic Statistics*, 38(2):243–256.

Christelis, D., Jappelli, T., and Padula, M. (2010). Cognitive abilities and portfolio choice. *European Economic Review*, 54(1):18–38.

Coe, N. B., von Gaudecker, H., Lindeboom, M., and Maurer, J. (2012). The Effect Of Retirement On Cognitive Functioning. *Health Economics*, 21(8):913–927.

Cornelissen, T. and Dustmann, C. (2019). Early School Exposure, Test Scores, and Noncognitive Outcomes. *American Economic Journal: Economic Policy*, 11(2):35–63.

Cornelissen, T., Dustmann, C., Raute, A., and Schönberg, U. (2018). Who Benefits from Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance. *Journal of Political Economy*, 126(6):2356–2409.

De Chaisemartin, C. and d'Haultfoeuille, X, X. (2020). Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, 110(9):2964–96.

De Chaisemartin, C. and D'HaultfŒuille, X. (2018). Fuzzy Differences-in-Differences. *The Review of Economic Studies*, 85(2):999–1028.

Dippel, C., Ferrara, A., and Heblich, S. (2020). Causal mediation analysis in instrumental-variables regressions. *Stata Journal*, 20(3):613–626.

Fratiglioni, L. and Wang, H.-X. (2007). Brain Reserve Hypothesis in Dementia. *Journal of Alzheimer's disease*, 12(1):11–22.

Frölich, M. and Huber, M. (2017). Direct and indirect treatment effects–causal chains and mediation analysis with instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1645–1666.

Garrouste, C. (2010). *100 Years of Educational Reforms in Europe: a contextual database.* PhD thesis, European Commission's Joint Research Centre (JRC).

Glymour, M. M., Kawachi, I., Jencks, C. S., and Berkman, L. F. (2008). Does childhood schooling affect old age memory or mental status? Using state schooling laws as natural experiments. *Journal of Epidemiology & Community Health*, 62(6):532–537.

Gorman, E. et al. (2017). Schooling, Occupation and Cognitive Function: Evidence from Compulsory Schooling Laws. *SocArXiv. October*, 23.

Grady, C. (2012). The cognitive neuroscience of ageing. *Nature Reviews Neuroscience*, 13(7):491–505.

Hampf, F. (2019). The effect of compulsory schooling on skills: Evidence from a reform in germany. Technical report, Ifo Working Paper.

Heckman, J. J. (2007). The Economics, Technology, and Neuroscience of Human Capability Formation. *Proceedings of the National Academy of Sciences*, 104(33):13250–13255.

Heckman, J. J. (2008). Schools, skills, and synapses. *Economic inquiry*, 46(3):289–324.

Heckman, J. J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic literature*, 48(2):356–398.

Heckman, J. J. and Vytlacil, E. (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica*, 73(3):669–738.

Heckman, J. J. and Vytlacil, E. J. (1999). Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects. *Proceedings of the National Academy of Sciences*, 96(8):4730–4734.

Hong, G. (2010). Ratio of mediator probability weighting for estimating natural direct and indirect effects. *Proceedings of the Biometrics Section – American Statistical Association*, pages 2401–2415.

Huber, M. (2014). Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics*, 29(6):920–943.

Huber, M. (2020). *Mediation Analysis*, pages 1–38. Springer International Publishing, Cham.

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71.

Imbens, G. W. and Rubin, D. B. (1997). Estimating Outcome Distributions for Compliers in Instrumental Variables Models. *The Review of Economic Studies*, 64(4):555–574.

Kamhöfer, D. A. and Schmitz, H. (2016). Reanalyzing zero returns to education in germany. *Journal of Applied Econometrics*, 31(5):912–919.

Kaufman, A. S. and Horn, J. L. (1996). Age changes on tests of fluid and crystallized ability for women and men on the kaufman adolescent and adult intelligence test (kait) at ages 17–94 years. *Archives of clinical neuropsychology*, 11(2):97–121.

Keele, L., Tingley, D., and Yamamoto, T. (2015). Identifying mechanisms behind policy interventions via causal mediation analysis. *Journal of Policy Analysis and Management*, 34(4):937–963.

Lindenberger, U. (2014). Human cognitive aging: Corriger la fortune? *Science*, 346(6209):572–578.

Lundborg, P., Nilsson, A., and Rooth, D.-O. (2014). Parental education and offspring outcomes: Evidence from the swedish compulsory school reform. *American Economic Journal: Applied Economics*, 6(1):253–78.

Mazzonna, F. and Peracchi, F. (2012). Ageing, cognitive abilities and retirement. *European Economic Review*, 56(4):691–710.

Mazzonna, F. and Peracchi, F. (2017). Unhealthy Retirement? *Journal of Human Resources*, 52(1):128–151.

Mogstad, M., Santos, A., and Torgovitsky, A. (2018). Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters. *Econometrica*, 86(5):1589–1619.

Nybom, M. (2017). The distribution of lifetime earnings returns to college. *Journal of Labor Economics*, 35(4):903–952.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 411–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ritchie, S. J. and Tucker-Drob, E. M. (2018). How Much Does Education Iimprove Intelligence? A Meta-analysis. *Psychological science*, 29(8):1358–1369.

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica*, 56(4):931–954.

Rohwedder, S. and Willis, R. J. (2010). Mental Retirement. *Journal of Economic Perspectives*, 24(1):119–38.

Salm, M., Siflinger, B., and Xie, M. (2021). The effect of retirement on mental health: Indirect treatment effects and causal mediation. Workingpaper, CentER, Center for Economic Research. CentER Discussion Paper Nr. 2021-012.

Schiele, V. and Schmitz, H. (2021). Understanding Cognitive-Decline in Older Ages: The Role of Health Shocks. *Ruhr Economic Papers 919, RWI*.

Schmitz, H. and Westphal, M. (2021). The dynamic and heterogenous effect of retirement on cognitive decline. *Ruhr Economic Papers 918, RWI*.

Schneeweis, N., Skirbekk, V., and Winter-Ebmer, R. (2014). Does education improve cognitive performance four decades after school completion? *Demography*, 51(2):619–643.

Seblova, D., Fischer, M., Fors, S., Johnell, K., Karlsson, M., Nilsson, T., Svensson, A. C., Lövdén, M., and Lager, A. (2021). Does Prolonged Education Causally Affect Dementia Risk When Adult Socioeconomic Status Is Not Altered? A Swedish Natural Experiment in 1.3 Million Individuals. *American Journal of Epidemiology*, 190(5):817–826.

Smith, J. P., McArdle, J. J., and Willis, R. (2010). Financial Decision Making and Cognition in a Family Context. *Economic Journal*, 120(548):363–380.

Sonnega, A., Faul, J. D., Ofstedal, M. B., Langa, K. M., Phillips, J. W. R., and Weir, D. R. (2014). Cohort Profile: the Health and Retirement Study (HRS). *International Journal of Epidemiology*, 43(2):576–585.

Steptoe, A., Breeze, E., Banks, J., and Nazroo (2003). Cohort profile: the English longitudinal study of ageing. *International Journal of Epidemiology*, 42(6):1640–1648.

Strittmatter, A., Sunde, U., and Zegners, D. (2020). Life cycle patterns of cognitive performance over the long run. *Proceedings of the National Academy of Sciences*, 117(44):27255–27261.

Sun, L. (2021). eventstudyinteract: interaction weighted estimator for event study.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.

Tchetgen, E. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *Annals of Statistics*, 40(3):1816–1845.

Tymula, A., Rosenberg Belmaker, L. A., Ruderman, L., Glimcher, P. W., and Levy, I. (2013). Like cognitive function, decision making across the life span shows profound age-related changes. *Proceedings of the National Academy of Sciences*, 110(42):17143–17148.

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

Westphal, M., Kamhöfer, D. A., and Schmitz, H. (2022). Marginal College Wage Premiums Under Selection Into Employment. *The Economic Journal*, forthcoming. ueac021.

Appendix: Additional tables and figures

Table A1: Number of observations

Country	Total	$CS=0$		$CS=1$		$D=0$		$D=1$	
		ERA$=0$	ERA$=1$	ERA$=0$	ERA$=1$	$M=0$	$M=1$	$M=0$	$M=1$
Austria	5705	473	2031	1548	1653	197	931	1868	2709
Germany	8273	2052	3332	1929	960	223	381	4179	3490
Spain	6764	2673	4091	0	0	1099	1401	2335	1929
Italy	2146	54	536	395	1161	78	207	1154	707
France	9799	1838	4445	2214	1302	547	1351	3958	3943
Greece	2419	1020	1399	0	0	431	281	1168	539
Czech Republic	3480	0	251	652	2577	2	12	1979	1487
England	42177	25827	16350	0	0	6409	10169	14081	11518
Total	80763	33937	32435	6738	7653	8986	14733	30722	26322

Table A2: Regression results: total treatment effect for years of education

	Treatment: More education			Treatment: Years of education		Reduced Form
	OLS (1)	First stage (2)	2SLS (3)	First stage (4)	2SLS (5)	(6)
More education (D)	1.423***		0.811***			
	(0.052)		(0.306)			
Years of education					0.997**	
					(0.457)	
Post CS-reform (Z_1)		0.244***		0.199**		0.198**
		(0.020)		(0.084)		(0.081)
Control variables	yes	yes	yes	yes	yes	yes

Number of observations in each regression: 80,763. Additional control variables are birth year fixed effects, interview wave fixed effects, country fixed effects, country-specific linear age trends, test repetition fixed effects and male. Standard errors in parentheses clustered on birth year-country level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Propensity Score ($P(Z_M)$) estimation

	(1)		(2)	
	Coefficient	Standard Error	Marginal Effect	Standard Error
Above ERA (Z_M)	0.285***	(17.71)	0.0869***	(17.84)
$D \times Z_D$				
1×1	-0.345***	(-14.94)	-0.105***	(-14.99)
0×1 (empty, no never takers)				
1×0	-0.258***	(-22.06)	-0.0788***	(-22.24)

Country				
Germany	-0.104	(-0.27)	-0.0318	(-0.27)
Spain	-0.193	(-0.46)	-0.0589	(-0.46)
Italy	-1.911**	(-2.66)	-0.583**	(-2.66)
France	-1.603***	(-4.21)	-0.489***	(-4.21)
Greece	1.919***	(4.13)	0.586***	(4.13)
Czech Republic	-5.177***	(-9.23)	-1.580***	(-9.25)
England	1.018**	(3.09)	0.311**	(3.09)
Country-specific age trends				
Germany	-0.0108	(-1.73)	-0.00329	(-1.73)
Spain	-0.00692	(-1.02)	-0.00211	(-1.02)
Italy	0.0237*	(1.98)	0.00724*	(1.98)
France	0.0220***	(3.48)	0.00673***	(3.48)
Greece	-0.0446***	(-5.76)	-0.0136***	(-5.76)
Czech Republic	0.0782***	(8.40)	0.0239***	(8.41)
England	-0.0289***	(-5.29)	-0.00881***	(-5.29)
Age	0.0393***	(3.54)	0.0120***	(3.54)
Birth year				
1931	0.396	(1.23)	0.121	(1.23)
1932	-0.0255	(-0.09)	-0.00777	(-0.09)
1933	0.376	(1.25)	0.115	(1.25)
1934	0.412	(1.38)	0.126	(1.38)
1935	0.0452	(0.16)	0.0138	(0.16)
1936	-0.0894	(-0.32)	-0.0273	(-0.32)
1937	-0.112	(-0.40)	-0.0341	(-0.40)
1938	-0.331	(-1.17)	-0.101	(-1.17)
1939	-0.438	(-1.53)	-0.134	(-1.53)
1940	-0.668*	(-2.31)	-0.204*	(-2.32)
1941	-0.762**	(-2.61)	-0.233**	(-2.61)
1942	-0.948**	(-3.21)	-0.289**	(-3.21)
1943	-1.045***	(-3.49)	-0.319***	(-3.49)
1944	-1.104***	(-3.64)	-0.337***	(-3.64)
1945	-1.318***	(-4.27)	-0.402***	(-4.28)
1946	-1.413***	(-4.51)	-0.431***	(-4.51)
1947	-1.487***	(-4.67)	-0.454***	(-4.68)
1948	-1.618***	(-5.00)	-0.494***	(-5.00)
1949	-1.741***	(-5.29)	-0.531***	(-5.29)
1950	-1.862***	(-5.56)	-0.568***	(-5.56)
1951	-1.917***	(-5.62)	-0.585***	(-5.62)
1952	-2.125***	(-6.12)	-0.649***	(-6.12)
1953	-2.191***	(-6.19)	-0.669***	(-6.19)
1954	-2.291***	(-6.35)	-0.699***	(-6.36)
1955	-2.469***	(-6.72)	-0.753***	(-6.72)
1956	-2.566***	(-6.86)	-0.783***	(-6.86)
Wave				
2	0.308***	(9.28)	0.0940***	(9.30)
3	0.402***	(7.85)	0.123***	(7.86)
4	0.576***	(8.43)	0.176***	(8.43)
5	0.749***	(8.49)	0.229***	(8.50)
6	0.891***	(8.23)	0.272***	(8.24)
7	1.292***	(10.01)	0.394***	(10.03)
8	1.218***	(7.98)	0.372***	(7.99)
Repetitions				

2	-0.190***	(-10.90)	-0.0581***	(-10.92)
3	-0.152***	(-7.74)	-0.0463***	(-7.75)
4	-0.181***	(-8.44)	-0.0553***	(-8.45)
5	-0.165***	(-6.26)	-0.0502***	(-6.26)
6	-0.0401	(-1.23)	-0.0122	(-1.23)
7	-0.137**	(-3.17)	-0.0418**	(-3.17)
8	0.147*	(2.52)	0.0450*	(2.52)
Male	-0.335***	(-30.73)	-0.102***	(-31.21)
Intercept	-0.336	(-0.40)		
N	76379	76379		

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A3: Regression results: Predicting Z_M with Z_D (conditional on controls)

	Dependent variable Z_M	
	Coefficient	Standard error
Compulsory schooling indicator Z_D:	0.026	(0.025)
Control variables	yes	

Number of observations: 80,763. Regression equation reads $Z_M = \alpha + \beta Z_D + \boldsymbol{X}'\boldsymbol{\delta} + \varepsilon$. Control variables are birth year fixed effects, interview wave fixed effects, country fixed effects, country-specific linear age trends, test repetition fixed effects and male. Standard errors in parentheses clustered on birth year-country level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.05$.

Figure A1: Potential outcomes in a randomized controlled trial

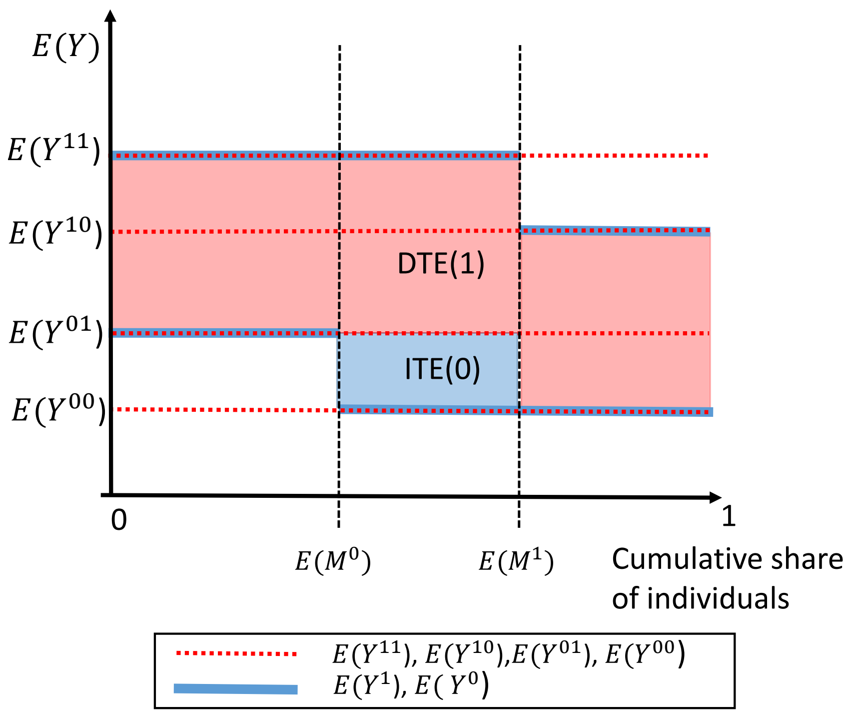
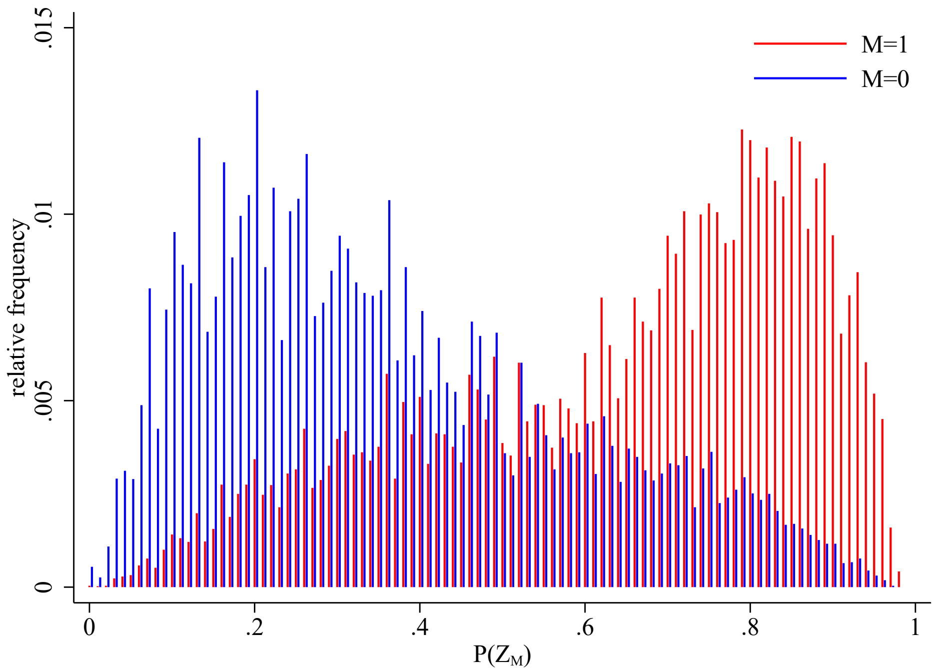
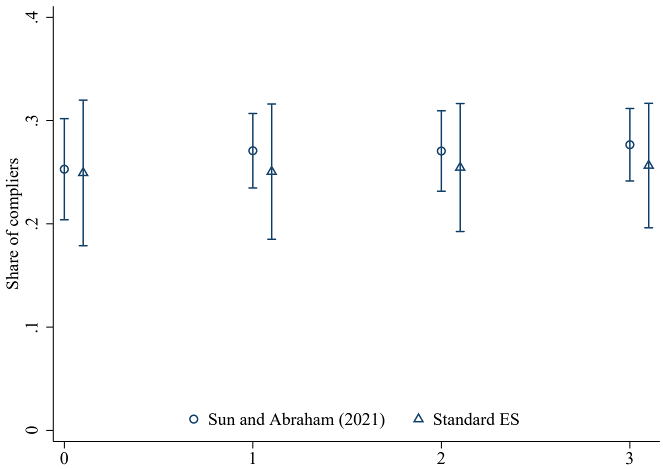


Figure A2: Support of $P(Z_M)$ by labor-force status (M)

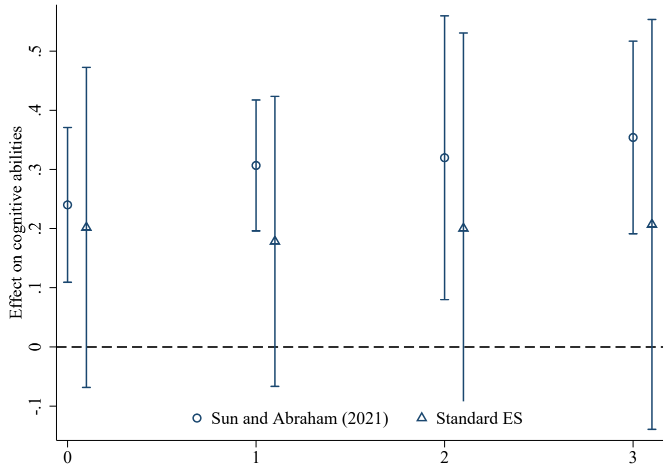


Notes: This graph plots the relative frequency of the $P(Z_M)$ values to lie in 0.01 bins of the propensity score by labor force status. The complete regression results are presented in Table A4.

Figure A3: Interaction-weighted event-study estimates (Sun and Abraham, 2021)



(a) First Stage: Compulsory schooling



(b) Reduced form: Compulsory schooling

Notes: This figure contrasts the conventional event-study results from Figure 9 with results from the interaction-weighted estimator suggested by (Sun and Abraham, 2021). Event time is birth cohort minus pivotal cohort. Here, we use the Stata command eventstudyinteract, see Sun (2021).

Early and Later-life Stimulation: How Retirement Shapes the Effect of Education on Old-age Cognitive Abilities

— For Online Publication: Supplementary Materials —

Hendrik Schmitz
Paderborn University,
RWI Essen, and
Leibniz Science Campus Ruhr

Matthias Westphal
TU Dortmund,
RWI Essen, and
Leibniz Science Campus Ruhr

September 2022

Hendrik Schmitz: Paderborn University, Warburger Strasse 100, 33098 Paderborn, Germany, Tel.: +49 5251 603213, E-mail: hendrik.schmitz@uni-paderborn.de.
Matthias Westphal: Technische Universität Dortmund, Fakultät Wirtschaftswissenschaften, 44221 Dortmund, Germany, Tel.: +49 231 755-5403,8, E-mail: matthias.westphal@tu-dortmund.de.

Supplementary Materials A: Retirement rules

If not further mentioned, early retirement eligibility (ERA) criteria are mainly based on Celidoni et al., 2017 and information of the website of the United States Social Security Administration (https://www.ssa.gov/policy/docs/progdesc/ssptw/, accessed in August 2022) as well as *Pensions at a Glance* by the OECD and MISSOC (Mutual Information System on Social Protection Comparative Tables Database. http://missoc.org,).

Austria
Men: ERA is 60 for birth cohorts until 1940 (September). It was stepwise increased to 65 until birth cohort 1952 (September).
Women: ERA is 55 for birth cohorts until 1945 (September). It was stepwise increased to 60 until birth cohort 1957 (September).

Czech Republic
Men: 57 for birth cohorts until 1952 and 60 for older cohorts. Number of contribution years increased from 25 to 35 between cohorts 1952 and 1959.
Women: ERA depends on birth year the number of children. It is 50 for birth cohorts until 1957 with 5 children and 54 for birth cohorts until 1954 with 0. For later birth cohorts, ERA is stepwise increased to 60 (reached for birth cohort 1959 with up to two children and a bit later for more than two children.

England: Early retirement age is the same as official retirement age.
Men: 65 for birth cohorts until 1953, 65 and 10 months for birth cohort 1954 and 66 for birth cohorts 1955 and younger.
Women: 60 for birth cohorts until 1949. Gradually increased to 66 for birth cohorts between 1950 and 1954. 66 for birth cohorts 1955 and younger.

France
Men and Women: ERA is 60 for birth cohorts until 1951 (June) and 62 for those born later.

Germany
Men: ERA is 63 with at least 15 contribution years.
Women: ERA is 60 until birth cohort 1951 and 63 thereafter for those with at least 15 contribution years.

Greece
Men: ERA is 62 with at least 15 contribution years.
Women: For women who started working before 1993: ERA is 55 with 15 contribution years. ERA is 50 for women with underage children and 18 contribution years. For women who started working since 1993: ERA is 60 with 15 contribution years. ERA is 50 for women with underage children and 20 contribution years.

Italy

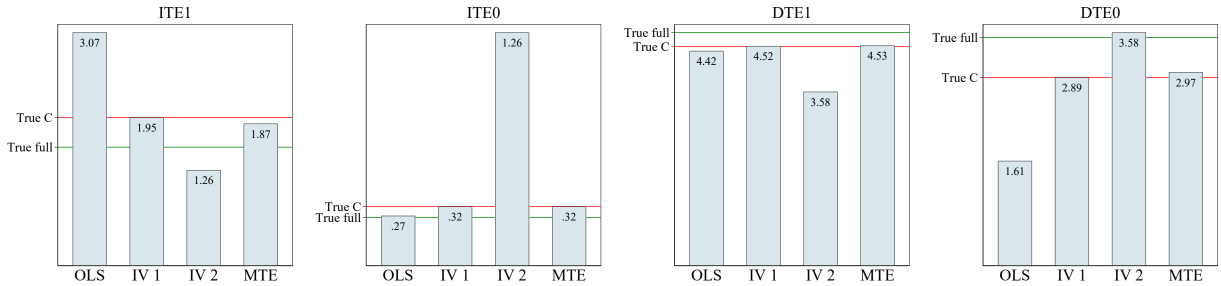
Men and Women: Between 1996 and 2012 stepwise increase from 52 (56 for self-employed) to 62. As of 2022: 64.

Spain

Men and women: Until birth cohort 1947: 63. Stepwise increase until birth cohort 1960 to 65.

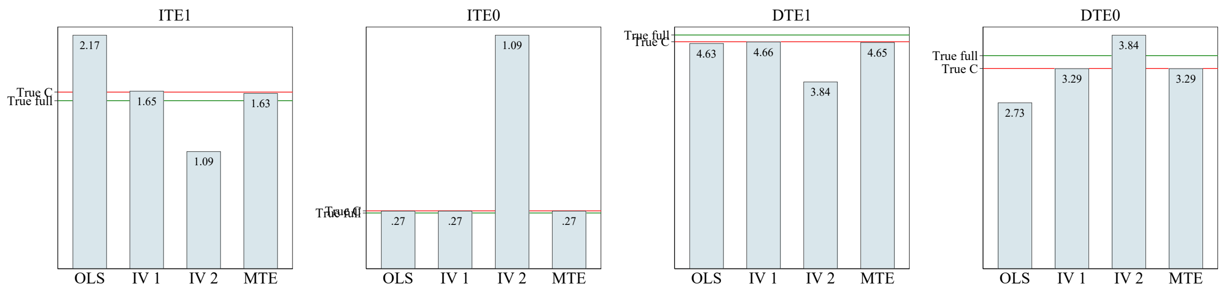
Supplementary Materials B: More simulation results

Figure S1: Results of DGP 1′



Notes: Own calculations. This Figure plots the average results of DGP 1′ after replicating the estimation respective procedures 200 times. Each panel refers to a specific indirect of direct mediation effect. The green and red horizontal line depict the average true effect for the whole sample and for the Z_1 compliers only (our target parameter), respectively. OLS and IV1 refer to implied mediation effects when $Y = \delta_0 + \delta_D D + \delta_M M + \delta_{DM} DM + \delta_X X + \epsilon$ and $M = \gamma_0 + \gamma_D D + \gamma_X X + u$ are estimated by OLS and two-stage least squares (2SLS), respectively. IV2 reports the implied effect when no interaction term is used in the 2SLS outcome regression. MTE refers to the results of the estimation procedure derived in the paper. Full results including RMSE are reported in Table S1.

Figure S2: Results of DGP 1″



Notes: Own calculations. This Figure plots the average results of DGP 1″ after replicating the estimation respective procedures 200 times. Each panel refers to a specific indirect of direct mediation effect. The green and red horizontal line depict the average true effect for the whole sample and for the Z_1 compliers only (our target parameter), respectively. OLS and IV1 refer to implied mediation effects when $Y = \delta_0 + \delta_D D + \delta_M M + \delta_{DM} DM + \delta_X X + \epsilon$ and $M = \gamma_0 + \gamma_D D + \gamma_X X + u$ are estimated by OLS and two-stage least squares (2SLS), respectively. IV2 reports the implied effect when no interaction term is used in the 2SLS outcome regression. MTE refers to the results of the estimation procedure derived in the paper. Full results including RMSE are reported in Table S1.

Table S1 reports the results using 200 rounds of simulation with 50,000 observations per round. The first two lines in each of the five blocks (denoted DGP 1, DGP 1′, DGP 1″, DGP 2, DGP 3) report the (unconditional) average total, indirect and direct effects, both for the full sample as well as for the subsample of compliers to instrument Z_1. We aim at estimating the effects for the compliers and compare average estimated effects to the average true effects.

Table S1: Simulation results

	TTE		ITE(1)		ITE(0)		DTE(1)		DTE(0)	
	$\widehat{E[\cdot]}$	RMSE	$\widehat{E[\cdot]}$	RMSE	$\widehat{E[\cdot]}$	RMSE	$\widehat{E[\cdot]}$	RMSE	$\widehat{E[\cdot]}$	RMSE
DGP 1										
True (full population)	5.06		1.56		.26		4.8		3.5	
True (compliers)	4.93		1.65		.27		4.66		3.29	
OLS	4.9	.75	2.17	9.08	.27	.18	4.63	.66	2.73	9.74
IV 1	4.94	.85	1.65	1.11	.27	.34	4.66	.81	3.29	.84
IV 2	4.94	.85	1.09	9.58	1.09	14.22	3.84	14.16	3.84	9.64
MTE	4.92	1.58	1.63	1.22	.27	.4	4.65	1.58	3.29	1.79
DGP 1′										
True (full population)	5.06		1.56		.26		4.8		3.5	
True (compliers)	4.84		1.95		.32		4.51		2.89	
OLS	4.69	2.65	3.07	19.51	.27	1	4.42	1.67	1.61	22.13
IV 1	4.84	.71	1.95	1.09	.32	.33	4.52	.69	2.89	.85
IV 2	4.84	.71	1.26	11.85	1.26	16.3	3.58	16.25	3.58	11.89
MTE	4.84	2.95	1.87	1.86	.32	.48	4.53	2.79	2.97	3.41
DGP 1″										
True (full population)	5.06		1.56		.26		4.8		3.5	
True (compliers)	4.93		1.64		.27		4.66		3.29	
OLS	4.89	.77	2.17	9.07	.27	.18	4.63	.66	2.73	9.73
IV 1	4.93	.89	1.65	1.14	.27	.34	4.66	.79	3.29	.74
IV 2	4.93	.89	1.09	9.59	1.09	14.19	3.84	14.15	3.84	9.61
MTE	4.92	1.64	1.63	1.26	.27	.38	4.65	1.58	3.29	1.61
DGP 2										
True (full population)	.98		.81		.57		.41		.16	
True (compliers)	1.44		.62		1.87		-.42		.82	
OLS	1.38	2.43	1.35	12.75	.14	29.93	1.24	28.89	.03	13.87
IV 1	1.44	4.19	.26	6.26	1.53	7.52	-.09	7.51	1.18	7.28
IV 2	1.44	4.19	.89	5.41	.89	17.13	.55	17.29	.55	6.13
MTE	1.48	4.67	.63	1.91	1.9	5.81	-.42	3.83	.85	3.42
DGP 3										
True (full population)	3.61		0.33		-0.97		4.58		3.29	
True (compliers)	8.32		-2.18		-2.36		10.68		10.50	
OLS	2.97	92.64	1.86	70.06	0.35	46.95	2.62	139.58	1.11	162.69
IV 1	8.16	4.26	0.27	42.50	0.99	58.16	7.19	60.66	7.91	45.08
IV 2	8.16	4.26	0.57	47.76	0.57	50.87	7.61	53.41	7.61	50.30
MTE	8.47	4.45	-2.29	3.54	2.23	4.02	10.70	4.39	10.76	6.39

Note: Simulation results on five different data generating processes (denoted DGP 1, DGP 1′, DGP 1″, DGP 2, DGP 3) described in the paper using 200 rounds of simulation with 50,000 observations per round. The columns refer to the treatment parameters TTE, ITE(1), ITE(0), DTE(1), and DTE(0) that we evaluate. For every data generating process, the first two lines refer to the true effects both for the full sample as well as for the subsample of compliers to instrument Z_1. The subsequent four lines refer to a separate estimator for the true effects. OLS and IV1 refer to implied mediation effects when $Y = \delta_0 + \delta_D D + \delta_M M + \delta_{DM} DM + \delta_X X + \epsilon$ and $M = \gamma_0 + \gamma_D D + \gamma_X X + u$ are estimated by OLS and two-stage least squares (2SLS), respectively. IV2 reports the implied effect when no interaction term is used in the 2SLS outcome regression. For all these estimators estimate the TTE by a OLS or 2SLS regression of Y on D. Finally, the MTE estimator refers to the estimator derived in the paper. These estimators are evaluated by the point estimate ($\widehat{E[\cdot]}$) and the root mean squared error (RMSE), which are presented in the columns. We aim at estimating the effects for the compliers and compare average estimated effects to the average true effects.

References

Celidoni, M., Bianco, C. D., and Weber, G. (2017). Retirement and cognitive decline. a longitudinal analysis using share data. *Journal of Health Economics*, 56:113–125.