Examining the Effect of Retirement on Cognitive Performance - A Unifying Approach *

János K. Divényi †
Central European University

September 28, 2020

Abstract

Several recent works investigate the effect of retirement on cognitive performance, arriving at different conclusions. The key ingredient of the various approaches is how they handle the endogeneity of the retirement decision. In order to examine this issue more deeply, I replicate the results of previous works using three waves from the Survey of Health, Ageing and Retirement in Europe (SHARE). I draw attention to potential biases inherent in the standard instrumental variable identification strategies and assess their magnitudes. Based on the lessons learned, I propose a new instrument that utilizes the panel structure of the data, enabling the comparison of individual cognitive paths. I show that if retirement has any adverse effect on cognitive performance it must be really small in magnitude.

*This paper is improved upon my Master thesis at Central European University. I thank Gábor Kézdi for advice, Gábor Békés, Zsombor Cseres-Gergely, Sergey Lychagin, Gábor Rappai, Anthony Strittmatter, Péter Zsohár and participants of the Hungarian Economics Association/Pécs University of Sciences, Faculty of Business and Economics' 2013 summer workshop for doctoral students and of the Venice International University Summer Institute on Ageing for helpful discussion and comments. The analysis was carried out in Stata with the help of the *ivreg2* package of Baum et al. (2014) and the tables were produced by the *estout* package of (Jann, 2005, 2007). The corresponding codes and the actual version of the paper can be found on GitHub. First draft: May 24, 2012.

†divenyi_janos@phd.ceu.edu

1 Introduction

In developed countries, increased life expectancy, together with the parallel decline in the average retirement age, has increased the average spell of retirement in the last decades (the expected number of years in retirement for OECD countries increased from 10.6 years in 1970 to 18.2 years in 2015 for men, and from 14.6 to 22.7 years for women[1]). Even if eligibility ages have been raised recently, people often spend 15-20 years of their lives as pensioners, which makes this phase of their life more and more relevant. Beside the individual level, the period of retirement is also of growing importance at the social level as well, because the proportion of retirees is increasing in the ageing population. As a natural consequence, various fields of research began to deal with the quality of the life of retirees. In this agenda, a particular aspect – namely the cognitive performance of old age individuals – has captured the attention of economists as it highly influences the decisions they make forming their consumption or saving behavior which affects the work of the economy to an increasing extent. Therefore, the age profile of cognitive abilities at the later stages of life is fundamental for many fields from marketing to pension and health policy.

It has been widely documented that individual cognitive performance tends to decline in older ages. According to Schaie (1989) cognitive abilities are relatively stable until the age of 50 but begin to decline afterwards. However, there is large heterogeneity in the progress of cognitive decay, raising the natural question of what are the driving forces behind and whether there is a way to decelerate it in order to maintain cognitive abilities as long as possible. A popular hypothesis, which is often called as use-it-or-lose-it hypothesis (see for example Rohwedder and Willis, 2010), suggests that the natural decay of cognitive abilities in older ages can be mitigated by intellectually engaging activities. Thus, retirement which goes together with the cease of cognitively demanding tasks at work, might accelerate the natural declining process, having a negative causal effect on cognition. In this respect, the notion of retirement simply refers to not working, and thus incorporates a broader definition than usual (for example, people on disability benefit or who are unemployed could also be regarded as retirees).

Many papers have been investigating recently the effect of retirement on cognitive abilities in developed countries (e.g. Rohwedder and Willis, 2010; Mazzonna and Peracchi, 2012; Bonsang et al., 2012), yet the results they have delivered are ambiguous. The inconclusive outcome is most likely due to the difficulty of identification and the resulting variety in the identification strategies.

In this paper I investigate the effect of retirement and cognition by two methods: First, I replicate the estimations of previous papers uncovering the factors behind the differences. I show that they fail to disentangle the true effect highlighting how sensitive

[1] OECD, https://stats.oecd.org/index.aspx?queryid=54758

their results are to minor modifications of the set of controls. Second, I apply a novel identification strategy which aims to handle the problems which the current literature suffers from. Applying a difference-in-differences approach I can account for all time-invariant individual heterogeneity, and get a smaller result than any other previous work. This result is robust to choosing different time-periods or including more controls.

My analysis is based on the first, second and the fourth waves of the Survey of Health, Ageing and Retirement in Europe (SHARE)[2] which collects rich multidisciplinary data about the socio-economic status, health (including cognitive functioning), and other relevant characteristics (like social networks) of people aged 50 or over across 10 developed European countries. The survey is harmonized not only across European countries but also with other surveys such as the Health and Retirement Study (HRS), that serves similar purposes in the United States and is used by some of the replicated papers (e.g. Bonsang et al., 2012). To my knowledge, this paper is the first which makes use of a large longitudinal cross-country sample to go after the effect of retirement on cognition.

The main challenge in uncovering the true effect is the endogeneity of retirement: a simple comparison of cognitive abilities of retirees and employees is likely to lead to biased estimates. Ideally, we would like to compare individuals from the same cohort and country, with the same age and education, one of them randomly assigned to be retired for a period of time while the other is working. As retirement is mainly an individual choice, this comparison is clearly impossible. For example, one can conveniently argue that the decay of cognitive abilities may induce the individual to retire, that is there is reverse causality going from cognition to retirement. This may result in overestimating the retirement effect on cognition in a simple comparison, even if we control for age. The standard solution is employing instrumental variables. Public policy rules (like official retirement age) seem to be good candidates for being relevant and exogenous instruments, as they clearly affect whether an individual retires but they generally refer to everyone irrespective of their actual cognitive performance.

While replicating previous cross-sectional analyses I show that applying public policy rule instruments in such settings can easily violate the exogeneity assumption: eligi-

[2]This paper uses data from SHARE wave 4 release 1.1.1, as of March 28th 2013(DOI: 10.6103/SHARE.w4.111) or SHARE wave 1 and 2 release 2.6.0, as of November 29 2013 (DOI: 10.6103/SHARE.w1.260 and 10.6103/SHARE.w2.260) or SHARELIFE release 1, as of November 24th 2010 (DOI: 10.6103/SHARE.w3.100). The SHARE data collection has been primarily funded by the European Commission through the 5th Framework Programme (project QLK6-CT- 2001-00360 in the thematic programme Quality of Life), through the 6th Framework Programme (projects SHARE- I3, RII-CT-2006-062193, COMPARE, CIT5- CT-2005-028857, and SHARELIFE, CIT4-CT-2006-028812) and through the 7th Framework Programme (SHARE-PREP, No 211909, SHARE-LEAP, No 227822 and SHARE M4, No 261982). Additional funding from 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 and OGHA 04-064) and the German Ministry of Education and Research as well as from various national sources is gratefully acknowledged (see www.share-project.org for a full list of funding institutions).

bility rules typically vary by country and gender, so not controlling for them properly can wrongly attribute general differences to the effect of retirement. This explains the sensitivity of cross-sectional results to alternative sets of controls.

Bonsang et al. (2012) use panel data from the HRS estimating a fixed-effect specification that naturally controls for such general differences. However, their results cannot be replicated on my sample. Also, their strategy uses an unbalanced panel over 10 years that mixes together the effect of retirement with other factors: cohort, attrition and learning effects could bias their results in an ambiguous way.

My strategy directly compares individual cognitive paths of employed and retired people. This way I also control for all time-invariant individual characteristics that allows for safely applying the standard public policy rules as instruments. Comparing cognitive paths implies a balanced panel specification, making it easier to assess panel concerns. I estimate the same relationship on different samples (comparing different waves) that contain different cohorts and also might suffer from different attrition rates, and my results are robust to these alternative specifications. Contrary to previous findings, my results suggest that retirement does not seem to cause serious harm for cognition: only around 0.02 standard deviation yearly.

This paper is structured as follows. Section 2 gives a formal setup for the problem and discusses the main challenges in the identification. Section 3 describes the data, detailing the various cognitive measures. Section 4 contains the replications of the previous results backed up by additional estimates that put them into context. Section 5 exhibits the results of my strategy. Section 6 concludes.

2 Model

A general way to model parametrically the relationship between cognitive performance and retirement is the following:

$$C_i = \alpha + f(R_i; \boldsymbol{\beta}) + u_i \tag{1}$$

$$u_i = \boldsymbol{X}_i' \gamma + \varepsilon_i \tag{2}$$

where C_i denotes the cognitive performance of individual i, R_i is the number of years the individual has spent in retirement (i.e. not working). I allow for the cognitive performance to depend upon these years through an arbitrary function f with parameter $\boldsymbol{\beta}$. The term u_i contains all factors associated with C_i except for R_i, for example: age. Equation (2) makes this dependency explicit where $\boldsymbol{X}_i$ is the vector containing these factors.

Clearly, $\mathrm{E}[C_i|R_i] = \alpha + f(R_i; \boldsymbol{\beta}) + \mathrm{E}[u_i|R_i]$. Assuming that we know f and have a good measure for C_i, the parameter of interest ($\boldsymbol{\beta}$) can be consistently estimated if $\mathrm{E}[u_i|R_i] = 0$. However, this is hardly the case. There are two sources which make the exogeneity assumption dubious: omitted variable bias and reverse causality.

Omitted variable bias There are lots of factors which are associated with the cognitive performance and also the years spent in retirement. These are factors in $\boldsymbol{X}_i$ which are correlated with R_i. The most obvious candidate is age: older individuals are expected to have spent more years in retirement and they also have worse cognitive skills due to age-related decline. Education is also incorporated in $\boldsymbol{X}_i$: worse educated individuals retire earlier and they also have worse cognition. One should take care of these factors when estimating the effect of retirement on cognitive performance. The main challenge here is that we do not know exactly what factors are in $\boldsymbol{X}_i$.

Reverse causality One can conveniently argue that the decay of cognitive abilities may induce the individual to retire, so there is reverse causality going from cognition to retirement. That may result in overestimating the retirement effect on cognition in a simple comparison, even if we control for all factors in $\boldsymbol{X}_i$.

Most attempts trying to uncover $\boldsymbol{\beta}$ apply instrumental variables, as they might be able to eliminate both problems. Good instrumental variables (let us denote them by the vector $\boldsymbol{Z}_i$) satisfy two requirements: first, they are correlated with the possibly endogenous retirement variable ($\mathrm{Cov}(\boldsymbol{Z}_i, R_i) \neq 0$), and second, they are related to the cognitive performance only through years of retirement ($\mathrm{E}[u_i|\boldsymbol{Z}_i] = 0$). If these two assumptions hold, both omitted variable bias and reverse causality are resolved.

3 Data

Most papers which are after the effect of interest use the same sources of data provided by three large longitudinal surveys: the Health and Retirement Study (HRS), the English Longitudinal Survey of Ageing (ELSA) and the Survey of Health, Ageing and Retirement (SHARE).

Aiming to provide a multidisciplinary data about ageing, the United States of America launched the Health and Retirement Study (HRS) in 1992, and since then the study has collected detailed information about socio-economic status, health (including cognitive functioning), and other relevant characteristics (like social networks) of people aged 50 or over. Respondents of the survey are visited biannually and put through in-depth interviews to collect rich panel micro data about ageing population. The English Longitudinal Survey of Ageing (ELSA) was designed according to the HRS with its first wave launched in 2002. 2 years later Continental Europe also decided to set up an ageing database by establishing the Survey of Health, Ageing and Retirement in Europe (SHARE), a cross-nationally comparable panel database of micro data. SHARE started with 12 countries (Austria, Belgium, Denmark, France, Germany, Greece, Israel, Italy, the Netherlands, Spain, Sweden and Switzerland) in 2004 with wave 1, three countries (the Czech Republic, Ireland and Poland) joined in wave 2, and another four countries (Estonia, Hungary, Portugal and Slovenia), joined in wave 4. The three surveys (HRS, ELSA and SHARE) are carefully harmonized, and thus provide an excellent basis for cross- country investigation of ageing population in developed countries.

What makes the surveys appropriate for this particular analysis is that they include a battery of tests about cognitive abilities (memory, verbal fluency and numeracy). The test of memory is done as follows: 10 simple words are read out by the interviewer and the respondent should recall them once immediately after hearing and then at the end of the cognitive functioning module. As a result, both immediate recall and delayed recall scores range from 0 to 10. Often, the two variables are merged to a composite one by adding them up, which is called total word recall. Verbal fluency is tested by asking the respondent to name as many distinct animals as she can within one minute. The length of this list provides a measure for verbal fluency. SHARE also consists of several questions about individual numeracy skills. Respondents who answer the first one correctly get a more difficult one, while those who failed get an easier one. The last question requires the respondent to calculate compound interest. The number of correct answers to these questions provides an objective measure of numeracy ranging from 0 to 4. Finally, there is a test of orientation of four questions which examines whether the respondent is aware of the date of the interview (day, month, year) and the day of the week. This test may be used to detect individuals with serious cognitive problems or progressed dementia.

Various measures of cognitive skills might grab its different aspects as argued in Mazzonna and Peracchi (2012). As most of the papers use the results on memory tests I also focus on that measure for comparison purposes. To have a common unit I use standardized scores to express scales in standard deviation.

Throughout the paper I make use of the first, second and fourth waves of SHARE. The third wave of data collection (SHARELIFE) is omitted, as it is of different nature: it focuses on people's life histories instead of current characteristics.

4 Replications

In this section I replicate the main results of the literature, specifically that of Rohwedder and Willis (2010), Mazzonna and Peracchi (2012), and Bonsang et al. (2012). I put all of these results in my unified framework and show that their differing conclusions actually fit in the broader picture. The ambiguity of their results stems from the differences in their identification strategies that implies that their estimated "effects" of retirement on cognitive performance measure different kinds of things.

The papers differ in three crucial aspect: first, what is their assumption about how retirement should affect cognitive performance (i.e. what is their assumption for f), second, how they handle omitted variable bias (i.e. which factors they are controlling for from $\boldsymbol{X}_i$), and third, what is their choice for instrumental variable to overcome endogeneity. Besides the methodology, they also differ in the data they use for estimation. However, considering the goal of uncovering a general relationship this fact should be of secondary importance as far as the measurements are comparable across the datasets.

The structural equation the papers try to estimate could be summarized as follows:

$$S_i = \alpha + f(R_i; \boldsymbol{\beta}) + \boldsymbol{X}_i^{*'} \boldsymbol{\gamma}^* + \tilde{u}_i \tag{3}$$

$$\tilde{u}_i = \tilde{\boldsymbol{X}}_i' \tilde{\boldsymbol{\gamma}} + \varepsilon_i \tag{4}$$

where S_i is a cognitive score, a measurement of cognitive performance. This formulation helps to differentiate between factors which are controlled for ($\boldsymbol{X}_i^*$) versus factors which remain in the error term ($\tilde{\boldsymbol{X}}_i$). To get a clear causal effect equation (3) is estimated by a 2SLS procedure where the first stage is

$$R_i = \boldsymbol{Z}_i' \boldsymbol{\pi} + \boldsymbol{X}_i^{*'} \boldsymbol{\rho} + \nu_i \tag{5}$$

From now on let us assume that the cognitive measurements detailed in the previous section describe well the actual cognitive skills. To be more precise, I assume that $C_i = S_i + e_i$ where e_i is a classical measurement error in the dependent variable, i.e. $\mathrm{Cov}(e_i, S_i) = \mathrm{Cov}(e_i, R_i) = 0$. In this case our estimators remain consistent although less precise.

All of the papers use various public policy rules to instrument retirement (such as pension eligibility rules). Such rules are good candidates for instrument as they vary across country and gender and are strongly correlated with employment status. The crucial question is whether it also satisfies the exogeneity assumption. Formally, the exogeneity assumption can be expressed as $\mathrm{E}\left[\tilde{u}_i | \boldsymbol{Z}_i\right] = 0$. It essentially says that there is no systematic difference in the cognitive performance of an eligible and a non-eligible individual in the sample (after controlled for some other features).

4.1 Rohwedder and Willis (2010)

The first serious attempt to uncover the causal relationship between retirement and individual cognitive performance (Rohwedder and Willis, 2010) uses a simple setup: they only include a dummy for not working on the right hand side on a restricted sample of people aged between 60 and 64. This is equivalent to estimating the average effect of retirement on cognition conditional on the average duration of retirement the sample, that is assuming that $f(R_i;\boldsymbol{\beta}) = \tilde{\beta}\mathbf{1}(R_i > 0)$ where $\tilde{\beta} = \beta\overline{R_i}$. Beside restricting the sample on a narrow age-range they do not include anything in $\boldsymbol{X}_i^*$. To handle endogeneity they use public pension eligibility rules as instruments: whether the individual is eligible for early or full benefits. See Table A1 for a summary of the methodologies.

Rohwedder and Willis (2010) estimate their model on the 2004 waves of SHARE, ELSA and HRS, and find that retirement has a large adverse effect on cognition among 60-64 years old, amounting to one-and-a-half standard deviation. Unfortunately, they do not report the average duration of retirement in their sample which makes it hard to convert this number to yearly average.

Using only the first wave of SHARE (and thus having a much smaller sample than theirs, 4464 versus 8828 observations) I was able to replicate their main findings (see the first column of Table 1). The pattern is the same: retirement seems to decrease cognitive performance. However, my estimation is somewhat smaller, amounting to only 1 standard deviation. Considering that the average duration of retirement in my sample is 6.6 years, it could be translated to an average yearly decline of 0.15 standard deviation (if I use the same number for conversion, the estimate of Rohwedder and Willis (2010) corresponds to 0.23 standard deviation yearly decline).

In order to be able to interpret the previous result as causal effect it should be true that $\mathrm{E}\left[\tilde{u}_i|\boldsymbol{Z}_i\right] = 0$. Clearly, eligibility rules are not related to unobserved individual idiosyncrasies in cognition, as they generally refer to everyone. So using the instrument indeed helps with the problems. However, there are other factors left in $\tilde{u}_i$ which are likely to be correlated with the instrument. For example, in most countries eligibility rules differ for males and females: women tend to become eligible earlier. Women also have higher memory scores than men in the same age, even before retirement (for people below 55 the mean difference amounts to 0.19 standard deviation in my data). Not controlling for gender is likely to lead to underestimated effects as women with better scores are overrepresented in the eligible population. Moreover, people from different countries might differ in their average education as well (e.g. because of different compulsory schooling laws affecting today's pensioners). As different countries also have different eligibility rules, ignoring schooling is also likely to undermine the

Table 1: Comparing the methodology of Rohwedder and Willis (2010) by two versions of the instrumental variable: 2SLS estimation

	(1) Rohwedder and Willis (2010)	(2) Mazzonna and Peracchi (2012)
Retired	−1.010***	−0.500***
	(0.14)	(0.13)
Constant	0.736***	0.365***
	(0.10)	(0.097)
Observations	4,464	4,464
Weak IV *F* statistic	154.45	156.12

Notes: Both results are from the second stage estimation of $S_i = \alpha + \beta \mathbf{1}(R_i > 0) + u_i$ where the retirement dummy is instrumented by early and normal eligibility dummies. The coefficient of interest in Rohwedder and Willis (2010) is –4.66*** on a sample of 8,828 observations which amounts to 1.5 standard deviation. The corresponding first stage regressions are summarized in Table A2.

Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

exogeneity of the instruments. Bingley and Martinello (2013) show that countries with higher eligibility ages also tend to have better educated old age people, and thus the effect of Rohwedder and Willis (2010) is overestimated. The violation of the exogeneity assumption makes the causal interpretation of the results in the first column of Table 1 questionable.

4.2 Mazzonna and Peracchi (2012)

The paper of Mazzonna and Peracchi (2012) improves upon Rohwedder and Willis (2010) along all of the three aspects: they allow for a yearly retirement effect instead of including just a retired dummy, they control for a set of features (age, gender and country), and they use a modified instrumental variable that has some variation within the country-gender cells. They end up with an estimated yearly decline of around 0.04 standard deviation, an order of magnitude less than the first estimate. I replicate their strategy by implementing their improvements one by one, to shed some light on what causes the reasonable drop in the effect.

I start with the modified instrumental variable: as opposed to the eligibility rules that were in effect at the time when the interviews were conducted, Mazzonna and Peracchi (2012) consider also the changes that the rules might have had during the times. For each individual they apply the eligibility rules that were in effect for the individual's cohort. This way they have some variation in the rules within country-gender cells. Both instrumental variables reach the same level of relevance (see the first stage regression results in Table A2 in the appendix). However, using the refined IV results in a reasonable drop in the coefficient of interest (see the second column of Table 1) even with the original specification. Introducing within-country-gender variation into the instrumental variable leads to halving of the effect, to a decline of only 0.075 standard deviation per year.

The methodology of Mazzonna and Peracchi (2012) differs from that of Rohwedder and Willis (2010) not only in respect of the instrumental variable. They also assume a different functional form, and control for a different set of features. Instead of using just a retirement dummy (and thus estimating the effect conditional on the average duration of retirement) they enter the number of years spent in retirement linearly in the equation (i.e. they assume that $f(R_i; \boldsymbol{\beta}) = \beta R_i$). To adapt to the different endogenous variable, they also modify the instrument accordingly: instead of using eligibility dummies, they calculate the years lived after reaching the eligibility age (i.e. $\max(0, \text{age} - \text{age}_{\text{eligibility}})$). They control for age in a different manner: instead of restricting the sample to 60-64 years old they estimate a linear age coefficient on a sample of people aged 50-70. They also control for country dummies and estimate the equation separately for men and women.

Table 2 summarizes the results of moving from the strategy of Rohwedder and Willis (2010) to that of Mazzonna and Peracchi (2012) step by step. (Table A3 in Appendix shows the corresponding first stage regression results.) This simple exercise illustrates how sensitive the estimates are to various specifications. In the followings, I comment on each specification, explaining what could be the reason behind the change (each point discusses the estimated specification with the corresponding number):

(1) The estimated effect of 0.05 standard deviation yearly (first column) is comparable to the effect estimated with the retirement dummy (see second column Table 1 and considering the average retirement duration of 6.6 years: 0.5 / 6.6 = 0.075).
(2) Extending the age range does not really matter.
(3) Restricting to those with labor market history makes the effect a bit larger. This is practically due to excluding some outliers with 50+ years spent in retirement.
(4) Controlling for age delivers weird results. The effect doubles and the coefficient on age is positive: age seems to improve cognitive performance until retirement, whereas it deteriorates it by around 0.14 standard deviation after that. This could be explained by country differences: as Bingley and Martinello (2013) draws the attention to, eligibility age and schooling is positively correlated (in my sample the correlation is 0.21 and 0.14 for the early and normal eligibility age, respectively). Therefore, comparing two individuals with the same age but differing years after eligibility likely means comparing two individuals from different countries with the older one being from the better educated country. This reasoning justifies the positive age coefficient and underlines the importance of controlling for both age and country.

Table 2: Moving from the strategy of Rohwedder and Willis (2010) to that of Mazzonna and Peracchi (2012)

	(1) aged 60-64	(2) aged 50-70	(3) + worked at 50	(4) + age	(5) + country
Years in retirement	−0.051***	−0.053***	−0.083***	−0.169***	0.158***
	(0.0072)	(0.0021)	(0.0029)	(0.015)	(0.032)
Age				0.044***	−0.112***
				(0.0075)	(0.015)
Constant	0.376***	0.359***	0.226***	−2.157***	6.090***
	(0.051)	(0.015)	(0.012)	(0.41)	(0.77)
Country dummies	No	No	No	No	Yes
Observations	4,052	17,448	14,052	14,052	14,052
Weak IV *F* statistic	118.05	1614.48	5779.83	246.16	60.57

Notes: The results are from the second stage estimation of $S_i = \alpha + \beta R_i + \boldsymbol{X}_i^{*'} \gamma^* + u_i$ with different samples and different $\boldsymbol{X}^*$ where years of retirement is instrumented by early and normal eligibility dummies. The corresponding first stage regressions are summarized in Table A3.
Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

(5) Controlling for country indeed solves the puzzle of positive age coefficient, changing its sign to what is expected. However, now the coefficient of interest changes sign and gets positive. The unexpected sign results again from omitted variable bias: gender is not controlled for. As mentioned previously, women perform significantly better on memory tests (controlling for age) than men (for this sample, they are by 0.28 standard deviation better. Thus, when we control for both age and country, we mainly identify the retirement effect from gender variation. To see that this is really the case, check the results in the first two columns of Table 3 where I also included a control for gender. The positive sign of the coefficient of interest reverses back to what is expected. The next two columns of the table shows the same result when the numeracy score is used to measure the cognitive skills. Women perform on average by -0.28 standard deviation worse on the numeracy test and correspondingly, we see larger negative effect of years in retirement on numeracy when not controlling for gender. For fluency, there is no notable difference in the performance of men and women.

There is one more puzzle in Table 3. Why is the coefficient on age is positive for numeracy when controlling for country but not for gender? (The same coefficient is negative for TWR.) There is a possible explanation for that: as the sample ages so decreases the share of women (interestingly, as mortality rates would predict the opposite). The coefficient on age is mainly identified on non-eligible population (as for eligible population the age effect is actually the sum of the coefficients on age and years in retirement). As women are better in memory tests, and their share is smaller in older cohorts, the com-

Table 3: Moving to the strategy of Mazzonna and Peracchi (2012) - the effect of gender control for different measures of cognitive performance

	(1) TWR	(2) TWR	(3) numeracy	(4) numeracy	(5) fluency	(6) fluency
Years in retirement	0.158***	−0.176***	−0.360***	−0.013	−0.038	−0.048
	(0.032)	(0.036)	(0.042)	(0.032)	(0.026)	(0.031)
Age	−0.112***	0.049***	0.151***	−0.016	−0.007	−0.002
	(0.015)	(0.017)	(0.020)	(0.016)	(0.013)	(0.015)
Female		0.290***		−0.302***		0.010
		(0.022)		(0.019)		(0.019)
Constant	6.090***	−2.270**	−7.318***	1.392*	0.951	0.689
	(0.77)	(0.90)	(1.03)	(0.80)	(0.64)	(0.76)
Country dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	14,052	14,052	14,145	14,145	14,004	14,004
Weak IV F statistic	60.57	43.41	61.11	44.23	62.38	45.15

Notes: The results are from the second stage estimation of $S_i = \alpha + \beta R_i + \mathbf{X}_i^{*'} \gamma^* + u_i$ where years in retirement is instrumented by early and normal eligibility dummies. Column (1) is equivalent to column (5) of Table 2, it is included to ease the comparison.

Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

position effect implies a negative coefficient for age. By contrast, the opposite is true for numeracy (women perform worse), so the composition effect implies a positive coefficient for age. Controlling for gender eliminates the level differences in cognitive scores. However, the rate of cognitive decline due to retirement might still be different by gender (i.e. heterogeneous retirement effect for men and women) that could further complicate the results and make the direction of possible bias hard to assess.

To allow for heterogeneous retirement effect by gender, Mazzonna and Peracchi (2012) estimate the equation separately for men and women in their preferred specification. Table 4 show my replication for their strategy for total word recall, numeracy and fluency. According to my results, the rate of decline is indeed different: the relatively better performing gender suffers a larger decline. These estimations differ from that of Mazzonna and Peracchi (2012) only in how cognitive performance is measured: they adjust the cognitive scores by the time spent on answering them whereas I do not do. Nevertheless, the replicated numbers are comparable to theirs, although less precise, showing a yearly decline of about 0.02-0.05 standard deviation.

Although this strategy is more robust than the original one, there is a potential issue that could contaminate the results. As already mentioned, education matters in old age cognitive skills even if gained in early stages of life (Banks and Mazzonna, 2012). Today's pensioners are highly affected by the expansion of average schooling: the 50 years old cohort spent on average 2.7 years more in school than the cohort of 70. Maz-

Table 4: Estimating separately by gender, closest to Mazzonna and Peracchi (2012)

	(1) TWR men	(2) TWR women	(3) numeracy men	(4) numeracy women	(5) fluency men	(6) fluency women
Years in retirement	0.015	–0.041*	–0.058	–0.035	–0.025	–0.029
	(0.037)	(0.024)	(0.039)	(0.024)	(0.037)	(0.022)
Age	–0.041**	–0.017	0.008	–0.009	–0.011	–0.015
	(0.017)	(0.012)	(0.018)	(0.012)	(0.017)	(0.011)
Constant	2.403***	1.309**	0.053	0.832	1.083	1.427**
	(0.90)	(0.61)	(0.94)	(0.61)	(0.89)	(0.56)
Country dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8,058	5,994	8,126	6,019	8,028	5,976
Weak IV F statistic	33.19	79.78	32.76	79.66	33.48	83.38

Notes: The results are from the second stage estimation of $S_i = \alpha + \beta R_i + \boldsymbol{X}_i^{*'}\gamma^* + u_i$ where years in retirement is instrumented. Mazzonna and Peracchi (2012) estimate different equations for immediate and delayed word recall that I use as an aggregate total word recall. Their estimates are –0.018** and 0.15* (men) and –0.051*** and –0.025*** (women) for IWR and DWR respectively, –0.029*** (men) and –0.041*** for numeracy, and –0.006 (men) and –0.023** (women) for fluency. See Column 2B of their Table 7.
Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

zonna and Peracchi (2012) try to control for education by including a low-education dummy (they also interact this dummy with the effect) and show that education indeed plays a significant role in explaining the heterogeneity in the levels of cognitive skills (and to a smaller extent in their age-related decline). However, there is also evidence (see for example the PISA surveys) that countries are different in how effectively they improve cognitive abilities in childhood. The first PISA survey was performed in 2000, the mean scores of countries in math, science and reading are positively correlated with the eligibility ages (OECD, 2001). If there is some persistence in the quality of education systems from the time when today's pensioners went to school and that of today, this might introduce a new type of bias in the estimations even if the number of years spent in education is controlled for.

Mazzonna and Peracchi (2012) improve a lot on the first estimate, getting a much smaller effect, but the sensibility of the results to different control sets show how hard it is to ensure the exogeneity of the instrument in a cross-sectional setting. As public policy rules only vary across countries and gender (mostly), and these are related to a lot of factors that also affect individuals' cognitive performance, controlling for all of them seems nearly impossible.

4.3 Bonsang et al. (2012)

The first paper that uses panel data to estimate the causal effect of retirement on cognitive performance (Bonsang et al., 2012) looks only at the US: they extract information from 6 waves (1998-2008) of the HRS.

In their main specification they follow the approach of Rohwedder and Willis (2010) by estimating the effect of retirement through a simple dummy, that is not taking into account the length of retirement. The only condition is that they restrict their attention to those who have been retired for at least one year. Formally, they assume that $f(R_i; \boldsymbol{\beta}) = \tilde{\beta}\mathbf{1}(R_i \geq 1)$.

Using unbalanced panel data they could control for individual heterogeneity by estimating a fixed effects specification. Thus, they could include a_i in $\boldsymbol{X}_i^*$ that means a control for all time-invariant factors, like (mostly) gender and country. Additionally, they also control for age in a quadratic form.

To handle the endogeneity of the retirement decision they also use eligibility ages as instruments: first, the age of 62 which is the eligibility age for social security in the US and second, the normal retirement age which varies by cohort. They measure the cognitive performance of the individual with the total word recall score. They find that being retired has an effect of 1 less word recalled, that could be translated into a yearly drop of 0.05 standard deviation, a slightly larger effect than what was found by Mazzonna and Peracchi (2012).

In the replication of their analysis, I use the same eligibility dummies for early and normal retirement benefits as before - these should correspond to the dummies which Bonsang et al. (2012) apply for the case of the US. Table 5 summarizes the results of the estimation which follows the main specification of Bonsang et al. (2012) but for a different sample (using SHARE instead of HRS). The age range (50-75) is the same, but the time period is shorter due to data limitation (spanning over only 6 years of 3 waves). The corresponding first stages can be found in the Appendix (Table A4).

I could not replicate their results. Both the unrestricted retirement dummy that was used by Rohwedder and Willis (2010) and the dummy requiring a retirement spell of at least one year lead to positive coefficient estimates. Nonetheless, none of them are significant.

The difference might originate from at least two sources: First, the US might be different from the European countries. Either because eligibility rules have a stronger influence on the time of retirement which is suggested by my weaker first stage, or because retiring might mean something else in terms of cognition. This is a topic worth investigating further. Second, my time period is shorter.

Table 5: Replication of Bonsang et al. (2012)

	(1) Retirement duration > 0	(2) Retirement duration ≥ 1
Retired	0.0955	0.173
	(0.14)	(0.14)
Age	0.185***	0.189***
	(0.016)	(0.017)
Age (sq.)	–0.00131***	–0.00135***
	(0.00013)	(0.00014)
Observations	41,476	37,374
Weak IV F statistic	136.82	131.89

Notes: The results are from the second stage estimation of $S_i = \alpha + a_i + \beta g(R_i) + u_i$ where $g(R_i) = \mathbf{1}(R_i > 0)$ or $\mathbf{1}(R_i \geq 1)$ and these dummies are instrumented by early and normal eligibility dummies. The coefficient of interest in Bonsang et al. (2012) is –0.942*** on a sample of 54,377 observations. This amounts to 0.27 standard deviation, or (considering the average duration of retirement in their sample) a yearly drop of 0.05 standard deviation. The corresponding first stage estimates are summarized in Table A4.

Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Estimating the fixed-effects model on an unbalanced panel makes the results hard to interpret. Mazzonna and Peracchi (2012) argue that panel data is inappropriate for identifying the causal effect of retirement on cognition because of two reasons: First, as in each wave the exact same cognitive exercises are performed, the participants may remember their answers from previous waves. Second, there is considerable attrition in the sample which might also bias the result. The unbalance in the data means that people who are present throughout the whole period get more weight. This makes it ambiguous how learning and attrition could bias the estimates. Also, pooling waves that extend over 10 years mixes cohorts that are differently affected by the extension of education. This could lead to compositional effect with unclear consequences. To assess the magnitude of these problems, I estimating the same relationship on different time periods. The resulting point estimates vary a lot, but all remain positive and insignificant (see Table A5 in the Appendix). We need a cleaner identification strategy to reliably estimate the causal effect.

5 My strategy

I demonstrated in the previous section that getting a clear causal effect in cross-section is really challenging. Some straightforward improvements upon the strategy of Rohwedder and Willis (2010) resulted in a drop of the estimated effect from 0.23 standard deviation per year to a magnitude less. But still, there are plenty of factors left in u_i^* which are likely to bias our results (to a yet unknown extent). Moving to panel identification solves a lot of issues. It enables us to compare the cognitive scores of the exact same individual instead of assuming that another one is a good subject for the comparison. The fixed-effects specification on unbalanced panel data resulted in a negative estimated yearly decline with comparable magnitude to what we get with the improved cross-sectional method, but this does not let itself be replicated on my data and might be subject to other biases resulting from attrition, learning and cohort effects.

I suggest a novel identification strategy that also exploits the panel nature of the data but aims for more clarity in the mechanisms. Instead of pooling periods, I look at changes between two periods to identify whether the cognitive paths differ for retired and non-retired individuals. Formally, I estimate the following equation

$$\Delta S_i = \alpha^* + \beta \Delta R_i + \boldsymbol{\theta}' \boldsymbol{W_i} + \Delta \tilde{u}_i \tag{6}$$

where ΔS_i is the change in cognitive performance between two waves (measured by TWR, numeracy or fluency score) and ΔR_i is the number of years spent in retirement during this period. Note that this specification implies a balanced panel. It also controls for all time-invariant individual heterogeneity like gender, amount and quality of education, country, etc.[3] Variables $\boldsymbol{W_i}$ control for different trends (instead of just level differences). This way I can handle issues like different rates of cognitive decline by gender, as discussed in the replication of Mazzonna and Peracchi (2012). The use of years spent in retirement instead of a simple retirement dummy captures better the actual treatment. To handle endogeneity, I use the same instruments as Mazzonna and Peracchi (2012): years after early and normal eligibility ages.

Following from the nature of the data collection, the time elapsed between interviews of two waves is not the same for all individual (e.g. it ranges from 11 to 40 months between the first two waves). Therefore, the amount of ageing between two waves is not the same across the sample so I need to control for this by including the years elapsed variable in $\boldsymbol{W}$.

This strategy makes evaluating the panel concerns easier. The learning effect is captured by the constant term: if people are indeed better at their second and third interviews because of the repetition, $\hat{\alpha}^*$ should be positive. However, learning effect shall not bias the coefficient of interest due to the balanced panel. It only matters if learning

[3]For two time periods, regression on differences gives the same coefficient estimates as the fixed effect specification on a balanced panel.

is different for employed and retired persons, but in this case, this is part of the retirement effect so it should be estimated within the treatment effect – and this is exactly what is going to happen in this specification.

Using a balanced panel also resolves some of the attrition concerns. To see whether it still causes any a problem, I will estimate the same relationship on different samples (comparing different waves) which are likely to suffer from different attrition rates and check whether they are different.

I have three waves of SHARE, so I can estimate the differences in three ways: between wave 1 and 4 (the longest period), between wave 1 and 2, and between wave 2 and 4. I restrict my sample to those who were aged between 50 and 70 and were either employed or retired in the first period, and has worked at age 50. I exclude those who returned to the labor market during this period (around 5% of the sample). Table 6 summarizes the data for the different comparison periods.

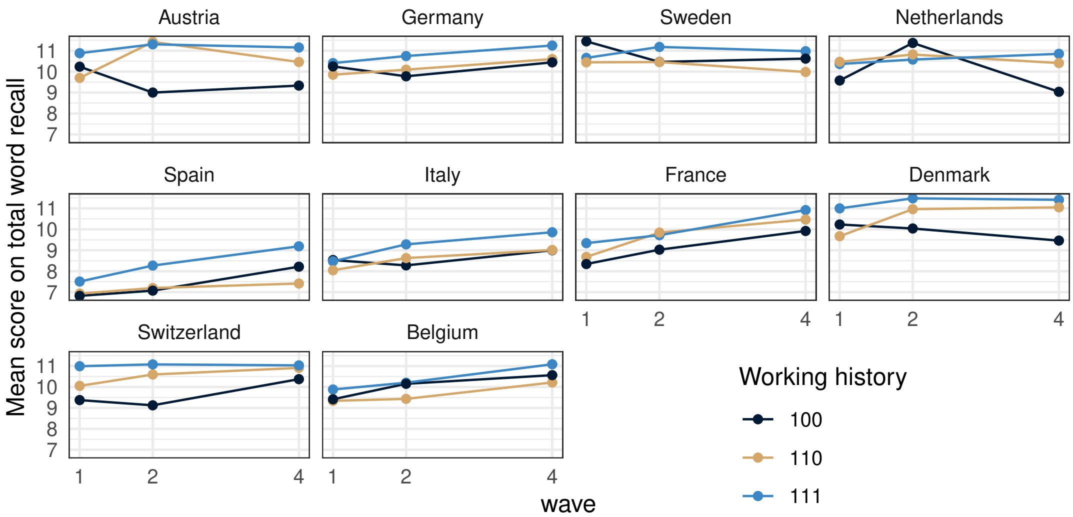
Table 6: Summary statistics

Sample		Men				Women			
		wave 1	wave 2	wave 4	N	wave 1	wave 2	wave 4	N
wave 1-4	age	59.7		66.3	3,551	58.9		65.5	2,877
		(5.9)		(6.0)		(5.8)		(5.9)	
	TWR	8.73		9.22	3,528	9.82		10.49	2,866
		(3.17)		(3.27)		(3.30)		(3.37)	
	numeracy	3.74		3.82	3,547	3.48		3.56	2,873
		(1.03)		(1.02)		(1.02)		(1.00)	
	fluency	21.34		20.41	3,510	22.00		21.25	2,858
		(7.08)		(6.94)		(7.04)		(7.24)	
	retired	1,537		2,244	3,551	1,137		1,761	2,877
wave 1-2	age	59.7	62.1		5,332	58.9	61.2		4,026
		(5.9)	(6.0)			(5.9)	(5.9)		
	TWR	8.71	9.01		5,258	9.78	10.13		3,998
		(3.13)	(3.17)			(3.26)	(3.23)		
	numeracy	3.76	3.80		5,295	3.48	3.54		4,000
		(1.01)	(1.04)			(1.04)	(1.03)		
	fluency	20.48	20.56		5,214	21.31	21.43		3,986
		(7.13)	(7.10)			(7.11)	(7.35)		
	retired	2,346	2,702		5,332	1,651	1,933		4,026
wave 2-4	age		60.3	64.5	4,359		59.2	63.5	3,783
			(5.8)	(5.8)			(5.8)	(5.9)	
	TWR		9.23	9.53	4,330		10.47	10.81	3,765
			(3.10)	(3.20)			(3.23)	(3.30)	
	numeracy		3.86	3.87	4,345		3.58	3.60	3,765
			(1.03)	(1.01)			(1.02)	(1.00)	
	fluency		21.69	20.87	4,313		21.31	21.43	3,757
			(7.17)	(6.94)			(7.14)	(7.07)	
	retired		1,712	2,142	4,359		1,319	1,711	3,783

Notes: The table presents the means and standard deviations (in parentheses) of the individuals' characteristics for each sample, by gender and wave.

Figure 1 shows the average paths of cognitive score of the individuals who were present in all waves, by working history for each country. For example, working history of 110 indicates a person who worked during the first two waves but left between wave 2 and 4. The first thing to note is that no clear pattern arises regarding the effect of retirement. There are some signs of selection and learning, but it does not seem like retirement would have any clear-cut effect on cognitive scores.

Figure 1: Pattern of cognitive scores across waves by working history



Notes: Working history code corresponds to the three waves used in the analysis, each digit displaying 1 if the individual worked in the given wave (100: worked only in first wave, 110: worked in first two waves, 111: worked in each wave).

The first three columns of Table 7 shows the 2SLS results for changes in total word recall score between wave 1 and 4 with different controls (again, years in retirement is instrumented by years after the eligibility ages). It seems that one more year in retirement decreases the cognitive score by 0.03-0.04 standard deviation point. My preferred specification is in column (3) where I allow for different trends in cognitive decline by gender and country. According to the estimates, women tend to lose their abilities slower. The positive constant term includes the effect of learning. The magnitude of the effect is comparable to that of Table 4, the closest replication of Mazzonna and Peracchi (2012). Note that my estimates are much more stable to including additional controls (e.g. allowing for different trends by country) which I interpret as a sign that the panel specification in itself eliminates many biases inherent in a cross-sectional analysis.

However, there is a reasonable scenario we have not considered yet. If natural age-related cognitive decline is concave instead of being linear then controlling only for a linear age trend might attribute the larger decline in older ages to the effect of retirement. Unfortunately, we are unable to allow for heterogeneous age effect and still use our instruments (there is not enough variability within country, gender and age). It still makes sense to run simple OLS regressions to see the difference.

Table 7: Panel estimation: change in total word recall score between wave 1 and 4

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	–0.025***	–0.026***	–0.040***	–0.028***	–0.016**
	(0.0058)	(0.0058)	(0.0060)	(0.0044)	(0.0065)
Years elapsed	–0.263***	–0.263***	–0.084*	–0.093*	–0.093*
	(0.033)	(0.033)	(0.050)	(0.050)	(0.050)
Female		0.042	0.053**	0.055**	0.049*
		(0.026)	(0.026)	(0.026)	(0.026)
Age at first wave					–0.009***
					(0.0032)
Constant	1.819***	1.808***	0.568*	0.571*	1.086***
	(0.22)	(0.22)	(0.34)	(0.34)	(0.39)
Country dummies	No	No	Yes	Yes	Yes
Observations	6,394	6,394	6,394	6,394	6,394
Weak IV F statistic	3729.06	3787.19	3665.10		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta\Delta R_i + \boldsymbol{W}_i^{*'}\boldsymbol{\nu} + \Delta\tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=1}$ and S_i = Total word recall$_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and $\boldsymbol{W}$ include years elapsed, female dummy and country dummies. For OLS, $\boldsymbol{W}$ additionally include age at first wave. The corresponding first stage estimates are summarized in Table A7.
Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

If we compare the third and fourth columns of the tables we can see that OLS results are slightly smaller in absolute value than those of 2SLS. Originally, we were afraid of a negative selection bias. This difference could be possibly explained by two facts: First, we eliminated all time-invariant individual heterogeneity so selection only matters if the rate of decline is different as well. Second, the 2SLS estimate the Local Average Treatment Effect (LATE), the effect of retirement on those who retired because reaching the eligibility age. These people might have been exposed a bigger change in their lifestyle than those who were retired anyway, and thus, might have been affected by a more expressed cognitive shock at retirement. Including age control (column 5) decreases the OLS estimate further, to less than 0.02 standard deviation yearly. This decline suggests that the 2SLS estimate in column (3) without age control slightly overestimates the true effect.

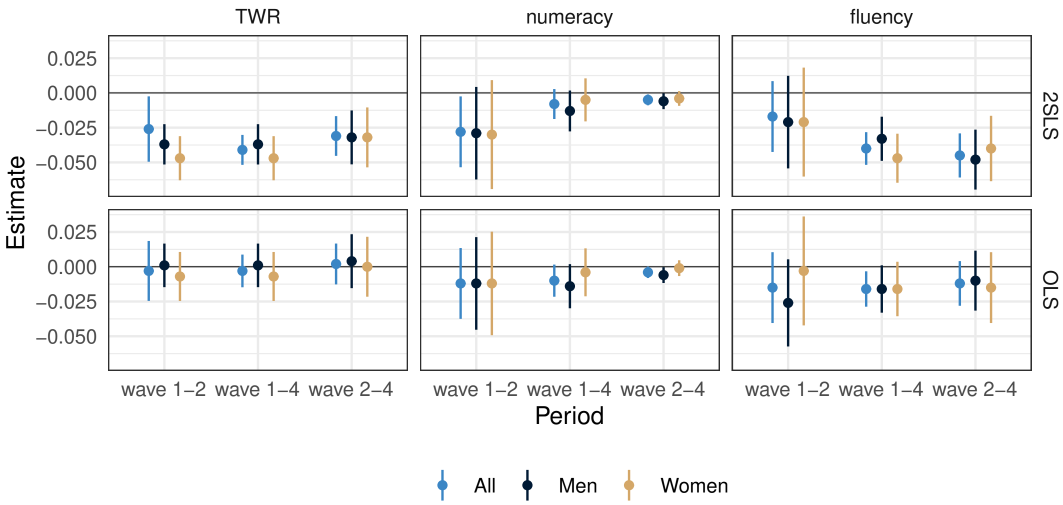
The same results for the other periods and other cognitive measures along with the corresponding first stages can be found in Appendix. The results are mainly consistent: 2SLS estimate a yearly decline of 0.01-0.05 standard deviation. The OLS estimates are smaller in absolute value, ranging between 0 and 0.02 standard deviation. There is no clear pattern by periods and measures which serves as a robustness check for the results[4]. My results are even smaller than the previous ones, like Mazzonna and Per-

[4]The other coefficients are less consistent. Sometimes, the constant term and the years elapsd estimate

acchi (2012) and Bonsang et al. (2012), especially if we take into account that the 2SLS estimates might still overestimate the true effect.

As an additional robustness check, we can estimate the preferred specifications separately for men and women. We saw before that is was crucial in the cross-sectional setting to pin down the estimates. In contrast, in the panel setting it is less of concern. Figure 2 compares the estimates by gender along with the pooled estimates. The estimated coefficients lie close to each other in every setting. The picture they draw is consistent: the effect of retirement on cognition cannot exceed a yearly 0.05 standard deviation decline (that is close in magnitude to what Mazzonna and Peracchi (2012) and Bonsang et al. (2012) get with different methods), and it might still be an overestimate of the true effect (see the near-zero OLS-estimates).

Figure 2: Comparison of the retirement effect estimate by gender



Notes: The dots represent $\hat{\beta}$, the lines show the 95% confidence intervals from the estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + \boldsymbol{W}_i^{*'} \boldsymbol{\nu} + \Delta \tilde{u}_i$, pooled and separately by gender. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and $\boldsymbol{W}$ include years elapsed, and country dummies. For OLS, $\boldsymbol{W}$ additionally include age at first wave.

6 Concluding remarks

Several recent works investigate the effect of retirement on cognitive performance, arriving to very different conclusions. To better understand the mechanisms that cause the excursive outcomes, I replicated the methods of the literature using data from the first three general waves of Survey of Health, Ageing and Retirement in Europe (SHARE). This exercise showed that identifying a clear causal effect is far from straightforward. Ideally, we would need to compare individuals from the same cohort and

switch sign – it can be a result of the high noise and the difficulty to separate the effects as there is a little variation in the years elapsed variable. The female coefficient is mostly positive.

country, with the same age and education, one of them randomly assigned to be retired for a period of time while the other being working. Such a comparison is clearly impossible. I traced back the major difference in the estimations to omitted variable bias: the large estimated retirement effect can be reduced a lot controlling for relevant variables such as gender and country. However, even after the inclusion of controls, the estimate is likely to be biased.

I came up with a novel identification strategy to solve the issues. I use the longitudinal feature of the data and compare the cognitive paths of employed and retired individuals. My estimated effect is even smaller than the smallest estimate in the literature, amounting to at most 0.05 standard deviation point decline yearly (that is still likely to be overestimated).

I reckon that retirement in itself has essentially no effect. What matters is the accompanying change in the lifestyle – that could be very heterogeneous across agents: some may retire from a stressful job to get more time to relax and care for grandchildren, others may lose the meaning of their lives. There are some works in the psychology literature that point to this direction. Stine-Morrow (2007) investigates the so called Dumbledore hypothesis, that individual choice plays a crucial role in old age cognitive decline. "it is our choices... that show what we truly are, far more than our abilities" (Rowling, 1999). Aichberger et al. (2010) find that physical inactivity goes together with worse cognitive performance. Hertzog et al. (2009) summarize previous results about what older adults can do to preserve their functional capacity. Future empirical work in this topic should build in these considerations and try to uncover the heterogeneous effect of retirement on cognitive performance, differentiating by the nature of the retirement, that is how it affected the individual's lifestyle.

Apart from the actual question I see my paper contributing to the literature in a more general way as well: it shows that replicating previous works and synthesizing their findings can lead to better understanding of the problem. Amidst of the replication crisis (Ioannidis, 2018) I consider this as a valuable lesson and hope to see more efforts towards this direction in the future.

References

Aichberger, M.C., M.a. Busch, F.M. Reischies, A. Ströhle, A. Heinz, and M.a. Rapp, "Effect of Physical Inactivity on Cognitive Performance after 2.5 Years of Follow-Up," *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, 2010, *23* (1), 7–15.

Angrist, Joshua D and Jörn-Steffen Pischke, *Mostly harmless econometrics: An empiricist's companion*, Princeton: Princeton University Press, 2008.

Banks, James and Fabrizio Mazzonna, "The Effect of Education on Old Age Cognitive Abilities: Evidence from a Regression Discontinuity Design.," *The Economic Journal*, 2012, *122* (560), 418–448.

Baum, Christopher F, Mark E Schaffer, and Steven Stillman, "IVREG2: Stata module for extended instrumental variables/2SLS and GMM estimation," 2014.

Bingley, Paul and Alessandro Martinello, "Mental retirement and schooling," *European Economic Review*, 2013, *63*, 292–298.

Bonsang, Eric, Stéphane Adam, and Sergio Perelman, "Does retirement affect cognitive functioning?," *Journal of Health Economics*, 2012, *31* (3), 490–501.

Hertzog, Christopher, Arthur F Kramer, Robert S Wilson, and Ulman Lindenberger, "Enrichment Effects on Adult Cognitive Development," *Psychological Science*, 2009, *9* (1), 1–65.

Ioannidis, John P.A., "Why most published research findings are false," *PLoS Medicine*, 2018, *2* (8), 2–8.

Jann, Ben, "Making regression tables from stored estimates," *The Stata Journal*, 2005, *5* (3), 288–308.

_ , "Making regression tables simplified," *The Stata Journal*, 2007, *7* (2), 227–244.

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

OECD, "Knowledge and Skills for Life," Technical Report 2001.

Rohwedder, Susann and Robert J Willis, "Mental Retirement.," *The Journal of Economic Perspectives*, 2010, *24* (1), 119–138.

Rowling, Joanne K, *Harry Potter and the Chamber of Secrets*, New York: Scholastic, Inc., 1999.

Schaie, K Warner, "The hazards of cognitive aging.," *The Gerontologist*, 1989, *29* (4), 484–93.

Stine-Morrow, Elizabeth A L, "The Dumbledore Hypothesis of Cognitive Aging," *Current Directions in Psychological Science*, dec 2007, *16* (6), 295–299.

Stock, James H, Jonathan H Wright, and Motohiro Yogo, "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic Statistics*, 2002, *20*, 518–529.

A Comparison of methodologies in the literature

Table A1: Comparing the methodologies of the literature

	Rohwedder and Willis (2010)	Mazzonna and Peracchi (2012)	Bonsang et al. (2012)
$f(R_i; \boldsymbol{\beta})$	$\tilde{\beta}\mathbf{1}(R_i > 0)$	βR_i	$\tilde{\beta}\mathbf{1}(R_i \geq 1)$
$\boldsymbol{X}_i^*$	-	age, gender, country dummies	age, age squared, individual fixed effects
$\boldsymbol{Z}_i$	normal & early eligibility dummies (no variation within country-gender cells)	normal & early eligibility dummies (some variation within country-gender cells)	social security eligibility & normal retirement age
used data	2004 waves of SHARE & HRS & ELSA	2004 wave of SHARE	6 waves (1998-2008) of HRS

Notes: Using a simple retirement dummy instead of years in retirement is equivalent to estimating the average effect conditional on the average time spent in retirement in the sample (or the average time excluding fresh retirees as in the case of Bonsang et al. (2012)), i.e. $\tilde{\beta} = \beta\overline{R_i}$.

B Additional tables for the replication exercises

This appendix contains summary tables of the first stage regressions for the replications exercises. It also contains a robustness check summary, replicating the method of Bonsang et al. (2012) on various subsamples.

Table A2: Comparing the methodology of Rohwedder and Willis (2010) by two versions of the instrumental variable: first stage

	(1) Rohwedder and Willis (2010)	(2) Mazzonna and Peracchi (2012)
Eligible for early benefits	0.323***	0.246***
	(0.028)	(0.027)
Eligible for full benefits	0.165***	0.185***
	(0.014)	(0.014)
Constant	0.375***	0.439***
	(0.027)	(0.025)
Observations	4,464	4,464
Adjusted R^2	0.0643	0.065

Notes: Both results are from the first stage estimation of $S_i = \alpha + \beta\mathbf{1}(R_i > 0) + u_i$ where the retirement dummy is instrumented by early and normal eligibility dummies. The corresponding coefficients for early and full benefits in Rohwedder and Willis (2010) are 0.19*** and 0.16**, respectively, with the adjusted R^2 being 0.059 on a sample of 8,828 observations.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A3: Moving from the strategy of Rohwedder and Willis (2010) to that of Mazzonna and Peracchi (2012): first stage

	(1) aged 60-64	(2) aged 50-70	(3) + worked at 50	(4) + age	(5) + country
Years after early eligibility	0.303***	0.016	0.183***	0.129***	0.033
	(0.071)	(0.033)	(0.013)	(0.013)	(0.025)
Years after normal eligibility	0.580***	0.581***	0.266***	0.144***	0.166***
	(0.079)	(0.033)	(0.013)	(0.015)	(0.018)
Age				0.200***	0.274***
				(0.013)	(0.024)
Constant	6.437***	8.310***	3.696***	−8.659***	−11.793***
	(0.38)	(0.18)	(0.071)	(0.83)	(1.42)
Country dummies	No	No	No	No	Yes
Observations	4,052	17,448	14,052	14,052	14,052
Adjusted R^2	0.0546	0.1561	0.4513	0.4599	0.4784

Notes: All results are from the first stage estimation of $S_i = \alpha + \beta R_i + \mathbf{X}_i^{*'}\gamma^* + u_i$ where the retirement dummy is instrumented by early and normal eligibility dummies.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A4: First stages, FE-IV estimation mimicing Bonsang et al. (2012)

	(1) Retired	(2) Retired for at least 1 year
Eligible for normal retirement	0.104***	0.110***
	(0.0072)	(0.0076)
Eligible for early retirement	0.0594***	0.0586***
	(0.0069)	(0.0070)
Age	0.000523	0.00491
	(0.0060)	(0.0059)
Age (sq.)	0.0000661	0.0000438
	(0.000046)	(0.000046)
Observations	41,476	37,374
Within-R^2	0.0575	0.0689

Notes: The results are from the first stage estimation of $S_i = \alpha + \beta g(R_i) + u_i$ where $g(R_i) = \mathbf{1}(YR_i > 0)$ or $\mathbf{1}(YR_i \geq 1)$ and these retirement dummies are instrumented by early and normal eligibility dummies. The corresponding coefficients for the eligibility dummies are 0.11*** and 0.07*** in Bonsang et al. (2012) on a sample of 54,377 observations with a within-R^2 of 0.242.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

C Detailed estimation tables for my strategy

This section contains the estimation results of my strategy for each cognitive score, for each period, along with the corresponding first stages. The main results of these tables are summarized in Figure 2 in Section 5.

Table A5: Replication of Bonsang et al. (2012) on various subsamples

	(1) wave 1-2	(2) wave 2-4	(3) wave 1-2-4
Retired for at least 1 year	0.394	0.248	0.127
	(0.65)	(0.19)	(0.15)
Age	0.105***	0.201***	0.202***
	(0.038)	(0.026)	(0.024)
Age (sq.)	−0.000587**	−0.00153***	−0.00145***
	(0.00029)	(0.00020)	(0.00018)
Observations	24,470	19,362	19,746
Weak IV F statistic	8.28	78.28	106.80

Notes: The results are from the second stage estimation of $S_i = \alpha + a_i + \beta g(R_i) + u_i$ where $g(R_i) = \mathbf{1}(R_i \geq 1)$ and this dummy is instrumented by early and normal eligibility dummies. The corresponding first stage estimates are summarized in Table A6.
Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A6: First stages, replication of Bonsang et al. (2012) on various subsamples

	(1) wave 1-2	(2) wave 2-4	(3) wave 1-2-4
Eligible for normal retirement	0.104***	0.110***	0.133***
	(0.0072)	(0.0076)	(0.0097)
Eligible for early retirement	0.0594***	0.0586***	0.0578***
	(0.0069)	(0.0070)	(0.0094)
Age	0.000523	0.00491	0.0272***
	(0.0060)	(0.0059)	(0.0087)
Age (sq.)	0.0000661	0.0000438	−0.000109
	(0.000046)	(0.000046)	(0.000067)
Observations	41,476	37,374	19,746

Notes: The results are from the first stage estimation of $S_i = \alpha + \beta g(R_i) + u_i$ where $g(R_i) = \mathbf{1}(YR_i \geq 1)$ and this dummy is instrumented by early and normal eligibility dummies.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A7: Panel estimation: change in total word recall score between wave 1 and 4: first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.270***	0.249***	0.215***
	(0.015)	(0.015)	(0.019)
Years after normal eligibility	0.071***	0.094***	0.132***
	(0.015)	(0.015)	(0.019)
Years elapsed	0.270***	0.255***	0.377***
	(0.067)	(0.066)	(0.096)
Female		−0.408***	−0.366***
		(0.052)	(0.051)
Constant	−0.594	−0.218	−0.041
	(0.47)	(0.47)	(0.67)
Country dummies	No	No	Yes
Observations	6,394	6,394	6,394

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + \boldsymbol{W}_i^{*'} \boldsymbol{\nu} + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=1}$, $S_i =$ Total word recall$_i$ and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A8: Panel estimation: change in total word recall score between wave 1 and 2

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	−0.019	−0.018	−0.017	−0.017*	−0.015
	(0.013)	(0.013)	(0.013)	(0.0090)	(0.013)
Years elapsed	0.001	0.002	0.085**	0.084**	0.084**
	(0.025)	(0.025)	(0.040)	(0.040)	(0.040)
Female		0.016	0.010	0.010	0.010
		(0.020)	(0.020)	(0.020)	(0.020)
Age at first wave					0.000
					(0.0024)
Constant	0.016	0.008	−0.094	−0.094	−0.067
	(0.059)	(0.060)	(0.11)	(0.11)	(0.18)
Country dummies	No	No	Yes	Yes	Yes
Observations	9,256	9,256	9,256	9,256	9,256
Weak IV F statistic	4272.83	4353.52	4332.50		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + \boldsymbol{W}_i^{*'} \boldsymbol{\nu} + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=2} - M_{i,w=1}$ and $S_i =$ Total word recall$_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and $\boldsymbol{W}$ include years elapsed, female dummy and country dummies. For OLS, $\boldsymbol{W}$ additionally include age at first wave. The corresponding first stage estimates are summarized in Table A9.
Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A9: Panel estimation: change in total word recall score between wave 1 and 2: first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.063***	0.052***	0.073***
	(0.0041)	(0.0042)	(0.0060)
Years after normal eligibility	0.061***	0.073***	0.054***
	(0.0041)	(0.0043)	(0.0059)
Years elapsed	0.388***	0.380***	0.374***
	(0.021)	(0.021)	(0.033)
Female		−0.173***	−0.146***
		(0.018)	(0.018)
Constant	0.070	0.226***	0.518***
	(0.056)	(0.058)	(0.096)
Country dummies	No	No	Yes
Observations	9,256	9,256	9,256

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta\Delta R_i + \boldsymbol{W}_i^{*'}\boldsymbol{\nu} + \Delta\tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=1}$, S_i = Total word recall$_i$ and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A10: Panel estimation: change in total word recall score between wave 2 and 4

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	−0.027***	−0.028***	−0.045***	−0.028***	−0.012
	(0.0079)	(0.0079)	(0.0081)	(0.0057)	(0.0082)
Years elapsed	0.040	0.041	−0.029	−0.043	−0.037
	(0.042)	(0.042)	(0.053)	(0.052)	(0.052)
Female		−0.003	0.008	0.010	0.003
		(0.023)	(0.023)	(0.023)	(0.023)
Age at second wave					−0.008***
					(0.0028)
Constant	−0.119	−0.119	0.011	0.020	0.463*
	(0.18)	(0.18)	(0.23)	(0.23)	(0.28)
Country dummies	No	No	Yes	Yes	Yes
Observations	8,095	8,095	8,095	8,095	8,095
Weak IV F statistic	4185.47	4205.43	4081.89		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta\Delta R_i + \boldsymbol{W}_i^{*'}\boldsymbol{\nu} + \Delta\tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=2}$ and S_i = Total word recall$_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and $\boldsymbol{W}$ include years elapsed, female dummy and country dummies. For OLS, $\boldsymbol{W}$ additionally include age at first wave. The corresponding first stage estimates are summarized in Table A11.

Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A11: Panel estimation: change in total word recall score between wave 2 and 4: first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.157***	0.147***	0.125***
	(0.0086)	(0.0087)	(0.012)
Years after normal eligibility	0.072***	0.083***	0.107***
	(0.0088)	(0.0090)	(0.012)
Years elapsed	−0.055	−0.048	0.234***
	(0.058)	(0.058)	(0.072)
Female		−0.189***	−0.168***
		(0.032)	(0.032)
Constant	1.310***	1.420***	0.863***
	(0.25)	(0.25)	(0.32)
Country dummies	No	No	Yes
Observations	8,095	8,095	8,095

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta\Delta R_i + \boldsymbol{W}_i^{*'}\boldsymbol{\nu} + \Delta\tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=2}$, S_i = Total word recall$_i$ and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A12: Panel estimation: change in numeracy score between wave 1 and 4

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	−0.008	−0.008	−0.008	−0.009**	−0.010
	(0.0053)	(0.0052)	(0.0055)	(0.0040)	(0.0059)
Years elapsed	−0.001	−0.001	0.048	0.049	0.049
	(0.029)	(0.029)	(0.045)	(0.045)	(0.045)
Female		−0.001	−0.002	−0.002	−0.002
		(0.023)	(0.023)	(0.023)	(0.024)
Age at first wave					0.000
					(0.0029)
Constant	0.035	0.035	−0.252	−0.252	−0.260
	(0.19)	(0.19)	(0.31)	(0.31)	(0.36)
Country dummies	No	No	Yes	Yes	Yes
Observations	6,420	6,420	6,420	6,420	6,420
Weak IV F statistic	3728.74	3787.09	3668.76		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta\Delta R_i + \boldsymbol{W}_i^{*'}\boldsymbol{\nu} + \Delta\tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=1}$ and S_i = Numeracy$_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and $\boldsymbol{W}$ include years elapsed, female dummy and country dummies. For OLS, $\boldsymbol{W}$ additionally include age at first wave. The corresponding first stage estimates are summarized in Table A13.

Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A13: Panel estimation: change in numeracy score between wave 1 and 4: first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.270***	0.250***	0.215***
	(0.015)	(0.015)	(0.019)
Years after normal eligibility	0.070***	0.093***	0.132***
	(0.015)	(0.015)	(0.019)
Years elapsed	0.275***	0.260***	0.382***
	(0.067)	(0.066)	(0.096)
Female		−0.409***	−0.366***
		(0.052)	(0.051)
Constant	−0.627	−0.250	−0.074
	(0.46)	(0.46)	(0.67)
Country dummies	No	No	Yes
Observations	6,420	6,420	6,420

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + \boldsymbol{W}_i^{*'} \boldsymbol{\nu} + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=1}$, $S_i = \text{Numeracy}_i$ and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A14: Panel estimation: change in numeracy score between wave 1 and 2

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	−0.026**	−0.028**	−0.028**	−0.020**	−0.012
	(0.013)	(0.013)	(0.013)	(0.0090)	(0.013)
Years elapsed	0.027	0.029	0.018	0.015	0.013
	(0.025)	(0.025)	(0.040)	(0.040)	(0.040)
Female		0.029	0.031	0.032	0.030
		(0.020)	(0.020)	(0.020)	(0.020)
Age at first wave					−0.002
					(0.0024)
Constant	−0.039	−0.053	0.023	0.019	0.154
	(0.059)	(0.060)	(0.11)	(0.11)	(0.18)
Country dummies	No	No	Yes	Yes	Yes
Observations	9,295	9,295	9,295	9,295	9,295
Weak IV F statistic	4282.96	4362.29	4341.05		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + \boldsymbol{W}_i^{*'} \boldsymbol{\nu} + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=2} - M_{i,w=1}$ and $S_i = \text{Numeracy}_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and $\boldsymbol{W}$ include years elapsed, female dummy and country dummies. For OLS, $\boldsymbol{W}$ additionally include age at first wave. The corresponding first stage estimates are summarized in Table A15.

Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A15: Panel estimation: change in numeracy score between wave 1 and 2: first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.063***	0.052***	0.073***
	(0.0041)	(0.0042)	(0.0060)
Years after normal eligibility	0.061***	0.073***	0.053***
	(0.0041)	(0.0043)	(0.0059)
Years elapsed	0.390***	0.382***	0.375***
	(0.021)	(0.021)	(0.033)
Female		−0.172***	−0.144***
		(0.018)	(0.018)
Constant	0.064	0.218***	0.515***
	(0.056)	(0.058)	(0.096)
Country dummies	No	No	Yes
Observations	9,295	9,295	9,295

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + \boldsymbol{W}_i^{*'} \boldsymbol{\nu} + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=2} - M_{i,w=1}$, S_i = Numeracy$_i$ and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A16: Panel estimation: change in numeracy score between wave 2 and 4

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	−0.003	−0.003	−0.005**	−0.005***	−0.004*
	(0.0020)	(0.0020)	(0.0020)	(0.0014)	(0.0021)
Years elapsed	0.030***	0.031***	0.018	0.018	0.018
	(0.010)	(0.010)	(0.013)	(0.013)	(0.013)
Female		−0.005	−0.006	−0.006	−0.006
		(0.0056)	(0.0056)	(0.0056)	(0.0057)
Age at second wave					0.000
					(0.00069)
Constant	−0.123***	−0.122***	−0.031	−0.031	−0.011
	(0.043)	(0.043)	(0.057)	(0.057)	(0.069)
Country dummies	No	No	Yes	Yes	Yes
Observations	8,110	8,110	8,110	8,110	8,110
Weak IV F statistic	4172.20	4190.63	4072.09		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + \boldsymbol{W}_i^{*'} \boldsymbol{\nu} + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=2}$ and S_i = Numeracy$_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and $\boldsymbol{W}$ include years elapsed, female dummy and country dummies. For OLS, $\boldsymbol{W}$ additionally include age at first wave. The corresponding first stage estimates are summarized in Table A17.

Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A17: Panel estimation: change in numeracy score between wave 2 and 4: first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.157***	0.147***	0.125***
	(0.0086)	(0.0087)	(0.012)
Years after normal eligibility	0.072***	0.083***	0.107***
	(0.0088)	(0.0090)	(0.012)
Years elapsed	−0.062	−0.055	0.231***
	(0.058)	(0.058)	(0.072)
Female		−0.184***	−0.164***
		(0.032)	(0.032)
Constant	1.338***	1.446***	0.872***
	(0.25)	(0.25)	(0.32)
Country dummies	No	No	Yes
Observations	8,110	8,110	8,110

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta\Delta R_i + \boldsymbol{W}_i^{*\prime}\boldsymbol{\nu} + \Delta\tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=2}$, $S_i = \text{Numeracy}_i$ and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A18: Panel estimation: change in fluency score between wave 1 and 4

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	−0.038***	−0.039***	−0.041***	−0.024***	−0.003
	(0.0053)	(0.0053)	(0.0055)	(0.0040)	(0.0060)
Years elapsed	0.008	0.008	0.058	0.046	0.046
	(0.030)	(0.030)	(0.046)	(0.046)	(0.046)
Female		0.018	0.042*	0.045*	0.036
		(0.024)	(0.024)	(0.024)	(0.024)
Age at first wave					−0.014***
					(0.0030)
Constant	0.076	0.071	−0.313	−0.311	0.540
	(0.20)	(0.20)	(0.31)	(0.31)	(0.36)
Country dummies	No	No	Yes	Yes	Yes
Observations	6,368	6,368	6,368	6,368	6,368
Weak IV F statistic	3733.22	3793.56	3670.54		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta\Delta R_i + \boldsymbol{W}_i^{*\prime}\boldsymbol{\nu} + \Delta\tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=1}$ and $S_i = \text{Fluency}_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and $\boldsymbol{W}$ include years elapsed, female dummy and country dummies. For OLS, $\boldsymbol{W}$ additionally include age at first wave. The corresponding first stage estimates are summarized in Table A19.

Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A19: Panel estimation: change in fluency score between wave 1 and 4: first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.270***	0.248***	0.213***
	(0.015)	(0.015)	(0.019)
Years after normal eligibility	0.072***	0.095***	0.135***
	(0.015)	(0.015)	(0.019)
Years elapsed	0.263***	0.247***	0.371***
	(0.067)	(0.067)	(0.097)
Female		−0.416***	−0.373***
		(0.052)	(0.051)
Constant	−0.541	−0.152	0.008
	(0.47)	(0.47)	(0.67)
Country dummies	No	No	Yes
Observations	6,368	6,368	6,368

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + \boldsymbol{W}_i^{*'} \boldsymbol{\nu} + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=1}$, $S_i = \text{Fluency}_i$ and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A20: Panel estimation: change in fluency score between wave 1 and 2

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	−0.021*	−0.021*	−0.026**	−0.015*	−0.003
	(0.011)	(0.011)	(0.012)	(0.0081)	(0.011)
Years elapsed	0.044*	0.044*	0.155***	0.150***	0.147***
	(0.023)	(0.023)	(0.036)	(0.036)	(0.036)
Female		0.005	0.011	0.011	0.009
		(0.018)	(0.018)	(0.018)	(0.019)
Age at first wave					−0.003
					(0.0022)
Constant	−0.084	−0.086	−0.587***	−0.594***	−0.404**
	(0.054)	(0.055)	(0.099)	(0.099)	(0.16)
Country dummies	No	No	Yes	Yes	Yes
Observations	9,200	9,200	9,200	9,200	9,200
Weak IV F statistic	4250.45	4330.57	4310.36		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + \boldsymbol{W}_i^{*'} \boldsymbol{\nu} + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=2} - M_{i,w=1}$ and $S_i = \text{Fluency}_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and $\boldsymbol{W}$ include years elapsed, female dummy and country dummies. For OLS, $\boldsymbol{W}$ additionally include age at first wave. The corresponding first stage estimates are summarized in Table A21.

Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A21: Panel estimation: change in fluency score between wave 1 and 2: first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.062***	0.051***	0.072***
	(0.0041)	(0.0042)	(0.0060)
Years after normal eligibility	0.062***	0.074***	0.055***
	(0.0042)	(0.0043)	(0.0060)
Years elapsed	0.388***	0.379***	0.375***
	(0.021)	(0.021)	(0.033)
Female		−0.173***	−0.145***
		(0.018)	(0.018)
Constant	0.074	0.230***	0.519***
	(0.056)	(0.058)	(0.096)
Country dummies	No	No	Yes
Observations	9,200	9,200	9,200

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta\Delta R_i + \boldsymbol{W}_i^{*'}\boldsymbol{\nu} + \Delta\tilde{u}_i$ where $\Delta M_i = M_{i,w=2} - M_{i,w=1}$, S_i = Fluency_i and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A22: Panel estimation: change in fluency score between wave 2 and 4

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	−0.036***	−0.036***	−0.031***	−0.016***	0.002
	(0.0072)	(0.0072)	(0.0073)	(0.0052)	(0.0075)
Years elapsed	−0.014	−0.015	0.192***	0.179***	0.185***
	(0.038)	(0.038)	(0.048)	(0.048)	(0.048)
Female		0.012	0.021	0.023	0.016
		(0.021)	(0.021)	(0.021)	(0.021)
Age at second wave					−0.008***
					(0.0025)
Constant	0.132	0.129	−0.693***	−0.686***	−0.222
	(0.16)	(0.16)	(0.21)	(0.21)	(0.25)
Country dummies	No	No	Yes	Yes	Yes
Observations	8,070	8,070	8,070	8,070	8,070
Weak IV F statistic	4173.04	4193.86	4070.78		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta\Delta R_i + \boldsymbol{W}_i^{*'}\boldsymbol{\nu} + \Delta\tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=2}$ and $S_i = \text{Fluency}_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and $\boldsymbol{W}$ include years elapsed, female dummy and country dummies. For OLS, $\boldsymbol{W}$ additionally include age at first wave. The corresponding first stage estimates are summarized in Table A23.

Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table A23: Panel estimation: change in fluency score between wave 2 and 4: first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.157***	0.146***	0.124***
	(0.0086)	(0.0087)	(0.012)
Years after normal eligibility	0.072***	0.084***	0.108***
	(0.0088)	(0.0090)	(0.012)
Years elapsed	−0.049	−0.042	0.240***
	(0.058)	(0.058)	(0.072)
Female		−0.192***	−0.172***
		(0.032)	(0.032)
Constant	1.289***	1.401***	0.848***
	(0.25)	(0.25)	(0.32)
Country dummies	No	No	Yes
Observations	8,070	8,070	8,070

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + \boldsymbol{W}_i^{*'} \boldsymbol{\nu} + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=2}$, $S_i = \text{Fluency}_i$ and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.