UNIVERSITÀ DEGLI STUDI DI PADOVA

Sede Amministrativa Università Degli Studi Di Padova

Dipartimento di Scienze Economiche ed Aziendali "Marco Fanno"

SCUOLA DI DOTTORATO DI RICERCA IN
ECONOMIA E MANAGEMENT

CICLO XXVI

Four Essays in Empirical Economics

Direttore della Scuola: Ch.mo Prof. Giorgio Brunello

Supervisore: Ch.mo Prof. Giorgio Brunello

Dottorando: Marco Bertoni

Contents

Acknowledgments

This thesis is the result of three years of intense, gripping and hard work, that I would have never been able to carry out without the invaluable support of many people, to whom I am indebted.

First of all, I wish to thank my supervisor, Giorgio Brunello, for his patience, his supportive presence, his skilful supervision and friendly advice, for the countless number of opportunities he provides me with and, most of all, for having taught me never to give up.

Lorenzo Rocco deserves special thanks as well: he always motivates and encourages me, and collaborating with him has offered me a unique opportunity of personal and professional growth.

I also owe big thanks to Erich Battistin, Danilo Cavapozzi, Luca Corazzini, Efrem Castelnuovo, Stefano Magrini, Andrea Moro, Luca Nunziata, Enrico Rettore, Elisabetta Trevisan, Christoph Weiss and, in particular, Guglielmo Weber.

My thanks also go to all the staff of the Centre for Economic Performance at London School of Economics, for the warm hospitality they have offered me during my visit, and in particular to Steve Machin and Sandra McNally, for having invited me to spend there an intense period of training.

During my stay at the CEP, I had the privilege to interact frequently with Olmo Silva, to whom I am grateful for his advice, that is always sincere and to the point, and for the trust he continuously puts in me. I also thank him and Steve Gibbons for having allowed me to substantially extend my stay at the CEP to work on our ongoing research project: I have benefitted immensely from this opportunity.

I thank all my PhD fellows in Padova and London, and in particular Alessio, Ambra, Andrea, Elena, Giulia, Felix, Luca, Marco, Marta, Michele, Monica, Paolo, Pietro, Thomas, and Richard, with whom I shared the pleasures and pains of this amazing experience.

I express profound gratitude to my mom, my dad and my brother, for their continuous support, advice, and love.

Finally, my warmest love and gratitude go to my soon-to-be wife Chiara, for her infinite love and for the sincere enthusiasm and empathy with whom she supports me in all my choices.

Introduction

This thesis is a collection of four essays in empirical economics.

The first chapter is titled "When the Cat Is Near, the Mice Won't Play: The Effect of External Examiners in Italian Schools". This paper is co-authored with Giorgio Brunello and Lorenzo Rocco. In this study, we use a natural experiment to show that the presence of an external examiner has both a direct and an indirect negative effect on the performance of monitored classes in standardized educational tests. The direct effect is the difference in the test performance between classes of the same school with and without external examiners. The indirect effect is the difference in performance between un-monitored classes in schools with an external examiner and un-monitored classes in schools without external monitoring. We find that the overall effect of having an external examiner in the class is to reduce the proportion of correct answers by 5.5 to 8.5% - depending on the grade and the test - with respect to classes in schools with no external monitor. The direct and indirect effects range between 4.3 and 6.6% and between 1.2 and 1.9% respectively. Using additional supporting evidence, we argue that the negative impact of the presence of an external examiner on measured test scores is due to reduced cheating (by students and/or teachers) rather than to the negative effects of anxiety or distraction from having a stranger in the class.

The second chapter is titled "Selection and the Age - Productivity Profile. Evidence from Chess Players", and is also co-authored with Giorgio Brunello and Lorenzo Rocco.We use data on professional chess tournaments to study how endogenous selection affects the relationship between age and mental productivity in a brain-intensive profession. We show that less talented players are more likely to drop out, and that the age-productivity gradient is heterogeneous by ability, making fixed effects estimators inconsistent. We correct for selection using an imputation procedure that repopulates the sample by applying to older cohorts the self-selection patterns observed in younger cohorts. We estimate the age-productivity profile on the repopulated sample using median regressions, and find that median productivity increases by close to 5 percent from initial age (15) to peak age (21), and declines substantially after the peak. At age 50, it is about 10 percent lower than at age 15. We compare profiles in the unadjusted and in the repopulated sample and show that failure to adequately address endogenous selection in the former leads to substantially over-estimating productivity at any age relative to initial age.

The third chapter is titled "Laterborns Don't Give Up. The Effects of Birth Order on Earnings in Europe", and is joint work with Giorgio Brunello. While it is well known that birth order affects educational attainment, less is known about its effects on earnings. Using data from eleven European countries for males born between 1935 and 1956, we show that firstborns enjoy on

average a 13.7 percent premium over laterborns in their wage at labour market entry. However, this advantage is short lived, and disappears by age 30, between 10 and 15 years after labour market entry. While firstborns start with a better match, partly because of their higher education, laterborns quickly catch up by switching earlier and more frequently to better paying jobs. We argue that a key factor driving our findings is that laterborns are more likely to engage in risky behaviours.

The fourth chapter is single-authored, and is titled " Hungry Today, Happy Tomorrow? Childhood Conditions and Self-Reported Wellbeing Later in Life". In this work, I use anchoring vignettes to show that, on data for eleven European countries, exposure to episodes of hunger in childhood leads people to adopt lower subjective reference points to evaluate satisfaction with life in adulthood. This is consistent with the satisfaction treadmill theory of hedonic adaptation, and highlights that failure to consider reporting heterogeneity will result in downward-biased estimates of the negative effects of starvation in childhood on the levels of wellbeing later in life. These findings underline the importance of considering issues of interpersonal comparability when studying the determinants of subjective wellbeing.

Introduzione

Questa tesi è una raccolta di quattro saggi in economia empirica.

Il primo capitolo è intitolato "*When the Cat Is Near, the Mice Won't Play: The Effect of External Examiners in Italian Schools*". L'articolo è co-autorato con Giorgio Brunello e Lorenzo Rocco. In questo studio utilizziamo un esperimento naturale per mostrare che la presenza di un osservatore esterno ha effetti negativi diretti ed indiretti sui risultati delle classi coinvolte in test didattici standardizzati. L'effetto diretto è dato dalla differenza nei risultati tra le classi della stessa scuola con e senza esaminatori esterni. L'effetto indiretto è invece calcolato come la differenza di risultati tra le classi non monitorate in scuole con un esaminatore esterno e le classi non monitorate in scuole senza controllo esterno. Dalle nostre stime emerge che la presenza di un osservatore esterno in una classe riduce la percentuale di risposte corrette in un *range* che va dal 5,5 all' 8,5% - a seconda del grado e della materia considerata - rispetto alle classi nelle scuole senza osservatori esterni. Gli effetti diretti e indiretti variano rispettivamente tra il 4,3 e il 6,6% e tra l'1,2 e l'1,9%. Utilizzando ulteriore evidenza empirica, concludiamo che l'impatto negativo della presenza di un esaminatore esterno sui punteggi ai test standardizzati è dovuto alla riduzione di comportamenti scorretti (il cosiddetto *cheating* - da parte di studenti e/o docenti), piuttosto che agli effetti negativi su ansia o distrazione dovuti alla presenza di un estraneo in classe.

Il secondo capitolo è intitolato "*Selection and the Age - Productivity Profile. Evidence from Chess Players*". Anche questo lavoro è co-autorato con Giorgio Brunello e Lorenzo Rocco. In questo studio, utilizziamo dati sui tornei di scacchi professionistici per studiare come la selezione endogena influenza la relazione tra età e produttività in una professione ad alta intensità cognitiva. Innanzitutto, mostriamo che gli scacchisti meno dotati sono più propensi ad abbandonare lo scacchismo agonistico nelle prime fasi della loro carriera, e che il gradiente età/produttività è eterogeneo in base all'abilità innata dei giocatori, rendendo gli stimatori ad effetti fissi non consistenti. Dunque, correggiamo per la selezione endogena utilizzando una procedura di imputazione che ripopola il campione applicando alle coorti più anziane i *pattern* di auto-selezione osservati nelle coorti più giovani, e stimiamo quindi il profilo età/produttività sul campione ripopolato attraverso regressioni sulla mediana. I nostri risultati evidenziano un aumento della produttività mediana di quasi il 5% dall'età iniziale (15 anni) al picco di età (21 anni), ed un declino sostanziale dopo il picco. A 50 anni, la produttività mediana è circa il 10% inferiore rispetto alla produttività dei quindicenni. Confrontando i profili stimati nel campione ripopolato e nel campione selezionato, concludiamo che non considerare la selezione endogena porta a sovrastimare sostanzialmente la produttività a qualsiasi età in relazione all'età iniziale.

Il terzo capitolo si intitola "*Laterborns Don't Give Up. The Effects of Birth Order on Earnings in Europe",* ed è co-autorato con Giorgio Brunello. Mentre è ben noto che l'ordine di nascita influenza il livello di istruzione, vi sono risultati meno chiari circa gli effetti dell'ordine di nascita sui redditi. Utilizzando dati per i maschi nati tra il 1935 e il 1956 in undici paesi europei, mostriamo che, rispetto ai loro fratelli, i primogeniti godono in media di un premio salariale del 13,7% all'ingresso nel mercato del lavoro. Tuttavia, questo vantaggio è di breve durata, e non è più presente già all'età di 30 anni, tra 10 e 15 anni dopo l'entrata nel mercato del lavoro. Mentre i primogeniti trovano inizialmente un lavoro di qualità migliore, in parte grazie alla loro maggiore istruzione, i secondogeniti colmano rapidamente queste differenze muovendosi prima e più frequentemente dei loro fratelli verso lavori meglio pagati. Nell'analisi, mostriamo che un fattore chiave per spiegare i nostri risultati riguarda la maggiore propensione dei secondogeniti ad assumere comportamenti rischiosi.

Il quarto capitolo è a firma singola, e si intitola "*Hungry Today, Happy Tomorrow? Childhood Conditions and Self-Reported Wellbeing Later in Life*". In questo lavoro vengono utilizzate le *anchoring vignette* per mostrare che, su dati relativi ad undici paesi europei, i soggetti esposti ad episodi di deprivazione nutrizionale nell'infanzia adottano standard di riferimento più bassi per valutare la propria qualità di vita nell' età adulta. Questo è coerente con la teoria del *satisfaction treadmill*, ed implica che la mancata considerazione di una possibile eterogeneità individuale negli stili di risposta porta a sottostimare gli effetti negativi dell'esperienza di deprivazione nutrizionale nell'infanzia sui livelli di benessere soggettivo nell'età adulta. Questi risultati sottolineano l'importanza di considerare problemi di comparabilità interpersonale nello studio delle determinanti del benessere soggettivo.

Chapter 1

When the Cat Is Near, the Mice Won't Play: The Effect of External Examiners in Italian Schools*

by

Marco Bertoni
(University of Padova and CEP, LSE)

Giorgio Brunello
(University of Padova, IZA and CEsifo)

Lorenzo Rocco
(University of Padova)

Abstract

We use a natural experiment to show that the presence of an external examiner has both a direct and an indirect negative effect on the performance of monitored classes in standardized educational tests. The direct effect is the difference in the test performance between classes of the same school with and without external examiners. The indirect effect is the difference in performance between un-monitored classes in schools with an external examiner and un-monitored classes in schools without external monitoring. We find that the overall effect of having an external examiner in the class is to reduce the proportion of correct answers by 5.5 to 8.5% - depending on the grade and the test - with respect to classes in schools with no external monitor. The direct and indirect effects range between 4.3 and 6.6% and between 1.2 and 1.9% respectively. Using additional supporting evidence, we argue that the negative impact of the presence of an external examiner on measured test scores is due to reduced cheating (by students and/or teachers) rather than to the negative effects of anxiety or distraction from having a stranger in the class.

Keywords: education, testing, external monitoring, indirect treatment effects.
JEL codes: C31, H52, I2.

* The authors are grateful to Erich Battistin, Thomas Breda, Daniele Checchi, David Figlio, Ifty Hussain, Edwin Leuven, Marco Manacorda, Guy Michaels, Hessel Oosterbeek, Steve Pischke, Olmo Silva and to the audiences at the 2012 LSE-CEP Annual Conference, the 2012 HECER Economics of Education Summer Meeting in Helsinki, the APPAM-INVALSI conference "Improving Education through Accountability and Evaluation" in Rome, and at seminars in Padova and CIDE (Bertinoro) for comments and suggestions. We also thank Patrizia Falzetti, Roberto Ricci and Paolo Sestito (INVALSI) for helping us with data collection and for explaining several technical features of the administration of the SNV tests. Financial support by the Ministry of Italian Universities (PRIN contract n. 2009MAATFS_002) is gratefully acknowledged. All errors are our own.

1. Introduction

A problem with test – based accountability systems in education is that they generate incentives for teachers, students and school administrators to "game" the system in order to obtain better results. The manipulation of test outcomes generates efficiency losses both when these outcomes are used to allocate resources to schools and teachers and when – more modestly – they provide valuable benchmarking information which can affect the choices of schools and their stakeholders.

One mechanism for inflating test scores is outright cheating. Empirical analysis of cheating behaviour is scarce[1]. In their influential study, Jacob and Levitt (2003) develop an algorithm for detecting teachers' cheating that combines information on unexpected test score fluctuations and suspicious patterns of answers for students in a class. They find that a small fraction of Chicago teachers responded to accountability pressures by completing student examinations in an attempt to improve outcomes.

A possible deterrent of forms of cheating that may occur during the test – e.g. students copying from one another or teachers communicating the correct answers – or during the scoring – e.g. teachers changing students' answers or filling in missing answers – is monitoring by external examiners. External monitoring has costs and benefits. Costs increase with the desired level of coverage. Benefits depend both on the efficiency gain associated to a reduction in cheating and on how effective monitoring is in influencing test scores and reducing cheating.

In this paper, we estimate the impact of external monitoring on test scores, using a rather unique natural experiment designed by the Italian central test administrator (INVALSI), which assigned external examiners to randomly selected classes and schools with the task of monitoring students taking the test and reporting results[2]. We compare test outcomes in the classes with an external examiner with the outcomes in other classes, where the test was administered by a local teacher, and find that the rate of correct answers is lower in the former than in the latter. Using additional supporting evidence, we argue that the negative impact of the presence of an external examiner on measured test scores is due to reduced cheating (by students and/or teachers) rather than to the negative effects of anxiety or distraction from having a stranger in the class.

Our study contributes to the literature on school accountability in two main directions. First, we show that the introduction of external examiners has a significant effect on measured test scores in an environment where there are incentives to manipulate results. Second, we document that the monitoring effects of having an external examiner spill over to un-monitored classes of the same school. We decompose the overall effect of external monitoring - which we measure as the

[1] See Figlio and Loeb, 2011, for a review of the recent literature.

[2] These tests are taken by the universe of primary second and fifth grade students. INVALSI sampled a number of classes and schools for external monitoring to obtain reliable data, speed up data collection and verification and prepare an annual report on the state of primary education in Italy.

difference in the average rate of correct answers in monitored classes and in classes of un-monitored schools - into a direct and an indirect effect. The direct effect is the difference in the test performance between classes with and without external examiners belonging to schools selected for external monitoring. The indirect effect is instead the difference in performance between un-monitored classes in a school with an external examiner and un-monitored classes in schools without external examiners.

We estimate that having an external examiner reduces the percentage of correct answers by 3.6 to 5.4 percentage points - depending on the grade and the test - which corresponds to 5.5 to 8.5% of the average score in classes belonging to schools with no external examiner. The estimated direct effect ranges from 2.8 to 4.2 percentage points (4.3 to 6.6%), and the residual indirect effect from 0.8 to 1.2 percentage points (1.2 to 1.9%). We discuss two alternative reasons why the effects of monitoring spread from the monitored class to the other classes in the same school. The first is that the presence of an external examiner in the school acts as a disciplinary device also on students and teachers in other classes of the same school because of the fear that the examiner may roam about. The second is that teachers dislike excessive dispersion in average class scores within the same school, because of the conflicts it could generate.

We find that the estimated overall effect of external supervision is significantly higher in the schools located in Southern Italy than in Northern schools and in schools where class size is smaller and the proportion of tenured teachers is higher. We show that territorial differences are associated to differences in social capital, even after controlling for territorial differences in GDP per capita and unemployment rates.

Studying the Italian experience with external monitoring has both advantages and disadvantages. The key advantage is that the random allocation of examiners to schools and classes allows us to bypass the selection problems that typically plague the evaluation of monitoring effects. A potential disadvantage is that in the Italian context there is limited accountability of schools and teachers. In this environment, the incentives to cheat may be weaker than in high-stakes contexts. In this case, our estimates can be interpreted as lower bounds of the effect of external monitoring in contexts where the incentives to manipulate results are stronger.

The paper is organized as follows: Section 2 reviews the relevant literature and Section 3 describes the design of the INVALSI test and the dataset. The empirical strategy is presented in Section 4. The main empirical results, a few robustness checks and extensions are reported in Section 5, 6 and 7, respectively. Conclusions follow.

2. Review of the literature

Aside from outright cheating studied by Jacob and Levitt (2003), the literature has identified several indirect ways that teachers and school administrators can use to manipulate student results. On the one hand, Jacob (2005), Figlio (2006), Figlio and Getzler (2006), Cullen and Reback (2006) and Hussain (2012) investigate whether schools engage in strategic manipulation of the composition of the pool of tested students by excluding low ability students, either by reclassifying them as disabled or by strategically using grade retention and disciplinary suspensions. On the other hand, Figlio and Winicki (2005) show that during testing periods some schools increase the caloric intake provided by school cafeterias so as to boost students' performance. Attempts to increase test scores by taking psycho-stimulant drugs are documented for the US by Bokhari and Schneider (2011), who show that the diagnosis of "attention deficit/hyperactivity disorder" is more frequent in states where there are stronger accountability laws.

To our knowledge, we are the first in this literature to investigate both the direct and the indirect effects of external examiners as deterrents of cheating in standardized tests. That indirect treatment effects can occur has been already pointed out by a broader literature. Heckman, Lalonde and Smith (1999), for instance, discuss how policy effects may spread to those not directly participating in the programme mainly because of general equilibrium or spill-over effects. Miguel and Kremer (2004) evaluate both direct and external effects of a Kenyan programme aimed at treating intestinal worms infection among primary school kids. In a similar fashion, Angelucci and De Giorgi (2009) evaluate the effects of *Progresa*, a Mexican aid programme based on cash transfers, and stress the importance of estimating indirect treatment effects on the ineligibles when there are social interactions between eligible and ineligible individuals.

3. The Design of INVALSI *Servizio Nazionale di Valutazione* (SNV) Tests and the Data

INVALSI[3] standardized tests in Italian and Math were introduced in Italian primary schools in 2008[4] to evaluate school productivity. The purposes of the evaluation[5] are to inform the central government about the general performance of the school system, and to offer schools a standardised reference to self-assess their strengths and weaknesses, using a value added approach. These tests are not formally high-stakes, because the allocation of resources to schools, the salary of

[3] INVALSI is the National Institute for the Evaluation of the Education System, in charge of the design and administration of standardized education tests in Italy.

[4] See Law n.147 – 2007, and Ministry of Education and Research Decree n.74 and 76 – 2009.

[5] See article 2 of the INVALSI statute (Ministry of Education and Research Decree n. 11-2011) and the Ministry of Education and Research Directive n. 88-2011.

teachers and the school career of students do not explicitly depend on test outcomes. Even so, pressure to perform well in the tests has been high because of the widespread expectations that they might be used at some point to evaluate teachers and schools. These expectations were fostered by the Ministry of Education, who in an intervention at the Lower House of the Italian Parliament (June 10th 2008) when the tests were introduced, made explicit reference to the need to establish within a few years a system of evaluation and incentives for teachers and schools based on student performance in the tests. Schools have an incentive to perform well also because results affect their reputation. Although the outcomes of the tests are not made public by INVALSI, schools have access to the results of their own students and can disclose them to parents and other stakeholders, in an effort to build their reputation and attract good students[6].

Since 2008 the tests have been administered every year. In this paper, we focus on the 2010 wave because of its peculiar design features. First, this wave was the first to test and collect data for the entire population of Italian primary school students in their second and fifth grade. Second, and most important for our purposes, in 2,000 randomly selected classes - out of a population of about 30,000 - the test was administered in the presence of an external examiner[7], who had two main tasks: a) be present in the class during the test and monitor its correct implementation; b) report student answers on the dedicated answer sheets and transmit them to INVALSI. In the other classes, the test was administered by teachers of the school (but not of the class and not in the subject tested), and reporting was done jointly with the teacher of the class. We use the random selection of classes as a natural experiment to estimate the effects of external monitoring on test outcomes.

Classes assigned to external monitoring were sampled using the same two-stage sampling scheme adopted by the IEA TIMSS survey, with stratification by region[8]. In the first stage, a pre-determined number of schools in each region were randomly selected by probabilistic sampling, with probability of inclusion proportional to school size, measured by the total number of students enrolled in the tested grades (second and fifth). In the second stage, and depending on school size, one or two classes for each tested grade within each treated school were selected by simple random sampling[9]. In each selected class, the test was administered in the presence of an external examiner. Table 1 shows for each grade the total and sampled number of primary schools, classes and pupils: about

[6] "INVALSI does not provide public rankings of schools based on the outcomes of the test. The main purpose of the tests is to provide each single school and its stakeholders with valuable information that can help them to benchmark and improve their performance. Each school is free to advertise its own results, using the tools provided by the Ministry of Education…" (free translation by the authors of Ricci and Sestito, 2012).

[7] External examiners were selected by INVALSI and the Regional Schooling Authorities mainly among retired teachers and active teachers employed in non-primary schools. Each examiner was paid 200 euro per working day. Details on the criteria adopted to select external examiners are reported in the Appendix.

[8] Region Valle d'Aosta and the Province of Bolzano autonomously decided to have all classes assigned to external monitoring. For this reason, we exclude them from our analysis. Our management of the data from the original to the final dataset is described in the Appendix.

[9] The average number of classes per school in sampled schools is 5.3, with a standard deviation of 1.9. Further details on the sampling procedure are reported in the Appendix.

18% of all primary schools and close to 7% of all classes and pupils in the second and fifth grade were selected to have an external examiner during the test.

We have access to the records containing the individual answers to the questions of the test taken in 2010 by students in classes with and without external monitoring, as they were transmitted to INVALSI by teachers and external examiners[10]. For each student, we also have information - provided by school offices - on her marks in Italian and Math during the semester before the test and on parental background. We add to these data the results of a questionnaire administered by INVALSI exclusively to fifth graders in order to collect additional information both on parental background and on student feelings and motivation during the tests. Finally, we have obtained from INVALSI additional information on school and class characteristics, including the number of students enrolled in each class and in each school for each tested grade, the proportion of tenured teachers in each school and, only for fifth grade students, an index of individual economic, social and cultural status (ESCS)[11].

We test for successful randomization by checking whether observables are balanced between sampled and non-sampled schools and classes. Reflecting the sampling strategy adopted by INVALSI, we verify balancing in two steps, first between sampled and non-sampled schools and second between sampled and non-sampled classes within the set of sampled schools. Since sampling is stratified by region and sampling probabilities depend on school size, our balancing tests are conditional on regional effects, school size and, in the second step, the number of classes in the school. Although we have data for second and fifth graders, we focus hereinafter on the latter for brevity. Selected results for second graders are shown in the Appendix.

For each variable X in Table 2 we first test between – school balancing by running

$$X_j = \alpha + \beta t_j + \rho RD_r + \sigma RS_{rj} + \varepsilon_j \qquad (1)$$

where the subscript r is for the region where the school is located, X_j is the average value of X in school j, t_j is a dummy taking the value 1 if school j has been sampled and 0 otherwise, RD_r is the full set of regional dummies, RS_{rj} is school size interacted with regional dummies and ε_j is the error term.

[10] All questions were either multiple choice or open questions with a univocally correct answer, and were coded by INVALSI as correct, incorrect or missing.

[11] Available information includes the following variables: 1) at the school level: whether the school offers a full time schedule; 2) at the class level: class size measured both as the number of students enrolled in the class and as the number of students who were present at the test, full or part-time schedule (measured in term of the schedule of the median student in the class, to avoid measurement errors); 3) at the individual level: gender, place of birth, citizenship, attendance of pre-primary school, age, employment, education and nationality of parents. For fifth grade students only we have information on: whether the student at home has own bedroom, internet access, an encyclopaedia, own desk, a computer and a place for doing homework, the number of books in the house, the number of siblings, whether she lives with both parents or not, the language spoken at home, whether she gets help with her homework or not.

Next, we test within-schools balancing by running

$$X_{ij} = \gamma + \delta t_{ij} + \varsigma R_{rj} + \upsilon_{ij} \quad (2)$$

where X_{ij} is the average value of X in the class i of school j, t_{ij} is a dummy that indicates whether class i in school j has been sampled and $R_{rj} = [RD_j, RS_{rj}, RC_{rj}]$ is a vector which includes the controls used in equation (1) as well as RC_{rj}, the number of fifth (or second) grade classes in school j interacted with regional dummies. We estimate equation (2) only for the classes belonging to the schools with external examiners and, since the second stage randomization took place within each school, we add school fixed effects and cluster standard errors at the school level.

Table 3 reports the point estimates of the β and δ coefficients in (1) and (2) and their statistical significance. Since balancing is not attained for the number of students enrolled in a class, which is greater among treated classes, we include this variable as a covariate in all our regressions. Turning to individual variables, although for some covariates we detect statistically significant differences across the various groups, the point estimates show that these differences are very close to zero in almost all cases. Prudentially, we add these variables as covariates in our regressions to eliminate the risk of unbalancing and to increase precision[12].

4. Identification and Estimation

We define the following three potential outcomes at the class level: Y_{00} if the class was assigned to a school with no external observer (an untreated class in an untreated school), Y_{11} in case of direct monitoring (a treated class in a treated school) and Y_{01} if the class was not monitored by an external examiner but belonged to a school where at least one other class was monitored (an untreated class in a treated school). By design, all classes of untreated schools are un-monitored.

Let the dummy variable S_j take the value one if school j has been assigned to school-level treatment (and zero otherwise) and the dummy C_i take value one if class i has been assigned to class-level treatment (and zero otherwise). The observed outcome Y_{ij} for class i in school j can be represented in terms of potential outcomes as follows:

$$Y_{ij} = (1 - S_j)Y_{00} + S_j C_i Y_{11} + S_j (1 - C_i) Y_{01} \quad (3)$$

[12] We notice that the proportion of missing data is slightly smaller in sampled schools and classes. This might be due to a more careful reporting of administrative information by secretaries in the schools and classes assigned to external monitoring.

We are interested in the identification and estimation of a) the average direct effect of monitoring $E[Y_{11}-Y_{01}]$; b) the average indirect effect of monitoring $E[Y_{01}-Y_{00}]$; c) the average overall effect of monitoring $E[Y_{11}-Y_{00}]$, where $E[.]$ is the mean operator.

The sampling procedure described in Section 3 is characterized by conditional randomization, which implies that a) in each region, the assignment to school - level treatment is random, conditional on the size of the school, measured by the number of students enrolled in the second and fifth grade; b) the assignment to class - level treatment for a class of a given grade in a treated school is random conditional on the size of the school, measured both by the number of students enrolled in the second and fifth grade and by the number of classes in the selected grade. Conditional randomization in each grade implies that

$$Y_{00},\ Y_{01},\ Y_{11} \perp S_j,\ C_i \mid R \tag{4}$$

When (4) holds, the average direct, indirect and overall effects of external monitoring are given by

$$E[Y_{11} - Y_{01} \mid R] = E[Y_{ij} \mid C_i = 1, S_j = 1, R] - E[Y_{ij} \mid C_i = 0, S_j = 1, R] \tag{5}$$

$$E[Y_{01} - Y_{00} \mid R] = E[Y_{ij} \mid C_i = 0, S_j = 1, R] - E[Y_{ij} \mid C_i = 0, S_j = 0, R] \tag{6}$$

$$E[Y_{11} - Y_{00} \mid R] = E[Y_{ij} \mid C_i = 1, S_j = 1, R] - E[Y_{ij} \mid C_i = 0, S_j = 0, R] \tag{7}$$

We aggregate our data at the class level and evaluate the effects of external monitoring on average class performance in the Math test by estimating

$$Y_{ij} = \theta_0 + \theta_1 C_{ij} S_j + \theta_2 S_j + \theta_3 R_{rj} + \theta_4 \Omega_{ij} + u_{ij} \tag{8}$$

where the dependent variable is the average percentage of correct answers in the class. We allow errors u to be correlated among the classes of the same school and weigh each class-level observation with the number of students in the class. The vector Ω includes for all grades the number of students enrolled in a class, which is greater among treated classes, and the following covariates: type of school (public or private), full or part-time schedule, average (in the class) gender, place of birth, citizenship, attendance of pre-primary school, age, grades in Italian and Math in the previous semester, employment, education and nationality of parents, and only for the fifth grade the percentage (in the class) of students who have their own bedroom, internet access, an encyclopaedia, own desk, a computer and a place for doing homework, the average number of books in the house, the average number of siblings, the percentage of students living with both parents, the language spoken at home, and whether they receive help with her homework. The

summary statistics of these covariates are in Panel A of Table 2. The direct, indirect and overall effect of external monitoring are given by θ_1, θ_2 and $\theta_1 + \theta_2$ respectively.

5. Results

Table 4 shows our baseline estimates of Eq. (8). Standard errors in this and the next tables are clustered at the school level. The first column in the table considers all Italian regions, and the remaining columns show the estimates by macro area (North, Centre and South). We find that having an external examiner in the class reduces the percentage of correct answers by 3.59 percentage points, which corresponds to a 5.5 percent decline with respect to the mean score in untreated schools[13]. Close to 80 percent (2.79/3.59) of this total effect is direct, and the remaining 20 percent (0.81/3.59) is indirect. The size of the total, direct and indirect effects varies with the macro area and is highest in Southern regions, where the total effect is -8.9%, and lowest in Northern Italy, where it is -2.6%.

Why are test results worse in classes with the external examiner? One possibility is that young students under-perform because they are distracted by the presence of a stranger in the class and are more anxious that students in un-monitored classes. The other possibility is that either students or teachers in classes without the external examiner engage in outright cheating[14]. We believe that the second one is the explanation, for the following reasons.

First, there is no evidence that students in classes with the external examiner are negatively affected in their feelings and motivation to complete the test properly. In a questionnaire filled up by fifth graders participating to the test in classes with and without the external examiner, INVALSI asked a set of motivational questions aimed at capturing the psychological status of students during the test, which included agreement or disagreement with the following sentences: a) I was already anxious before starting the test; b) I was so nervous I couldn't find the answers; c) while answering, I felt like I was doing badly; d) while answering, I was calm. Table 5 presents the results of estimating Eq. (8) when the dependent variable is the percentage of students in the class agreeing with each of the four statements above. We find no evidence that being in a class with an external examiner increased

[13] As shown in Table A.2 in the Appendix, the total effect is somewhat larger for second graders (5.4 percentage points, or 8.5% of the average score in untreated schools).

[14] We assume that cheating is unlikely in classes with the external examiner. On the one hand, since schools are informed of having been selected to receive an external examiner only about one week prior to the date of the test, there is little room of *manoeuvre* for teachers to react and adopt strategies that manipulate student performances in the presence of the examiner. On the other hand, we assume that external examiners have no incentive to cheat and collude with school teachers and principals in order to boost school results. In support of this assumption, INVALSI (2010a) used a procedure to detect cheating in monitored classes and concluded that there was no evidence of cheating. The cheating detection algorithm is described in INVALSI (2010b).

anxiety or nervousness. Quite the opposite, there is evidence that students in these classes were less nervous and calmer during the test.

Second, we examine the distribution of results within classes. In the absence of external controls, the teacher can communicate the correct answers to students or change their answers in the answer sheet, or students can simply copy from each other. If outright cheating by students and/or teachers was taking place in the classes without the external examiner, we should find that in these classes – *ceteris paribus* - the standard deviation and the coefficient of variation of test results are lower than in classes with the external examiner, where cheating is minimized or altogether absent. While distraction and anxiety can reduce average performance, it is not obvious that they reduce its variability. Table 6 shows for the entire country the effects of the presence of an external examiner on the within – class standard deviation and coefficient of variation of the percentage of correct answers, as well as on the bottom quartile, median and top quartile of the distribution of test scores within classes.

We find that in classes with the external examiner the standard deviation and the coefficient of variation of results are about 6% and 11% higher than in un-monitored classes. There is also evidence that the presence of the external examiner affects to a higher extent the performance of students in the lower quartile of the distribution of outcomes, in line with the expectation that cheating typically helps low performers, or that low performing students are those more prone to copy. When compared with students in untreated schools, having an external examiner reduces the score of these students by about 8% (-4.26/55.6). This effect is strongest for second grade students in Southern Italy, where if reaches a striking 18.7%[15].

Third, we compute an index of heterogeneity in the pattern of answers given by students in each class. For each question, we use a modified version of the Herfindahl Index

$$H = \frac{1 - \sum_{a=1}^{A} s_a^2}{1 - \frac{1}{A}}. \qquad (12)$$

where s_a is the within-class share of students who chose answer "a" in the set A of possible answers[16]. Index H ranges between 0 and 1, with higher values signalling a more heterogeneous pattern of answers to a given question. We obtain an overall measure of the heterogeneity of answers in the class by averaging H across all questions in the test. While we expect this measure to decline in classes without the external examiner in the presence of cheating, it is not clear whether it

[15] Detailed results by macro area are available from the authors upon request.

[16] We treat missing values as a separate category.

declines or increases if anxiety or distraction play a role. Table 7 reports the estimates of Eq. (8) when the dependent variable is *H*, and shows that heterogeneity is significantly higher in classes with the external examiner. We also find that, as in the case of the percentage of correct answers in the class, the effects of external monitoring on the heterogeneity of answers increase significantly moving from Northern to Southern Italy (columns (2) to (4)).

Finally, the correlation between test score outcomes and teacher grades in the semester before the tests should be lower in the presence of cheating. Using individual student data, we examine the correlation between the rank in the test and the rank in teacher grades in classes with and without the external examiners. In line with our expectations, we find a higher correlation for students taking the test in classes with the external examiner[17].

While these results are suggestive of the presence of cheating, we cannot say whether cheating occurs because teachers change answers in their report to INVALSI, or because they suggest the correct answers to students in the class, or because students are given extra time or are allowed to copy from each other in classes without the external examiner. Since all these cheating strategies generate a higher proportion of correct answers and a lower within - class dispersion of results, they are observationally equivalent in our data. To distinguish between some of these strategies, we would need to observe both the answers directly chosen by students and the answers reported by teachers to INVALSI. Unfortunately, we only observe the latter. We can only speculate that since in un-monitored classes teachers are responsible for supervision in class, collection of the tests, filling-in of the answer sheets on the basis of the responses given by the students and transmission of the answer sheets to INVALSI, they have certainly plenty of opportunities to modify test results.

An interesting and novel result of our analysis is that external examiners affect performance not only in the class they supervise but also in other classes of the same school. This indirect effect of monitoring in school tests has not been detected before and deserves further explanation. One interpretation is that teachers administering the test in the same school where the external examiner is present are afraid to be monitored by this supervisor and therefore restrain their cheating activities. This interpretation relies on irrational behaviour, because teachers were informed before the test that the external examiner's mandate was restricted to the randomly selected class.

An alternative explanation is that teachers dislike excessive dispersion in average test scores within the same school, because such dispersion could generate conflicts with other teachers. To illustrate, consider a school where a single class is supervised by an external examiner. If teachers administering the test in the other classes cheat freely, these classes will look much better than the supervised class, where cheating is restrained. This may generate conflicts with the teacher in charge

[17] We regress the individual within-class rank in the test on the individual within-class rank in teacher grades and its interaction with the presence of an external examiner and find that the interaction attracts a positive and statistically significant coefficient, especially in the South, where cheating appears to be more widespread. Detailed results are available from the authors upon request.

of the supervised class. To reduce these conflicts, teachers in un-monitored classes may be induced to restrain their cheating.

6. Robustness checks

In this section we investigate whether our main results are robust to several sensitivity checks. First, since the dependent variable of our main estimates is a fraction (the percentage of correct answers in the class) we implement the GLM estimator proposed by Papke and Wooldridge (1996) to deal with fractional dependent variables. Estimated marginal effects, shown in Table A.4 in the Appendix, are in line with the baseline estimates in Table 4.

Second, we exploit the census nature of our data and the fact that we observe almost the entire population of students in each grade to apply a finite population correction to statistical inference. Results (Table A.5 in the Appendix) are qualitatively unchanged with respect to the baseline, but precision increases significantly.

Third, we drop all observable covariates not required for the implementation of conditional randomization[18]. Since assignment to treatment does not depend on observables, finding differences between the estimates that include and exclude covariates is a symptom of strategic manipulation of the composition of the pool of tested students. Results in Table A.6 in the Appendix do not provide any strong evidence in this direction. Finally, we test for differences in absenteeism across treatment statuses, using as dependent variable the percentage of students absent from the test in each class. Again, differences in behaviour across the three groups are minimal (see Table A.7 in the Appendix).

7. Extensions

So far, we have allowed treatment effects to vary across the different macro areas of the country. Yet there might be other relevant sources of heterogeneity to be considered. In this section we do two things. We start by exploring what these other sources could be – without pretending to be exhaustive - and then examine whether regional heterogeneity is related to regional differences in social capital.

Our candidate sources of heterogeneous treatment effects are a) class size; b) the percentage of tenured teachers in the school; c) an indicator of average parental background for the students in the class[19]. On the one hand, if student cheating is easier in larger classes, we should find that the overall effect of having an external examiner increases with class size. On the other hand, larger classes

[18] We still include regional dummies, regional dummies interacted with school size and with the number of fifth grade classes in the school, and the number of students enrolled in the class.

[19] Descriptive statistics for these variables are shown in Table 2 – Panel B.

could increase the cost of cheating by teachers or could reduce the effectiveness of external supervision. In this case, the overall effect should be smaller in larger classes. Column (1) in Table 8 presents our estimates when both the direct and the indirect effect are interacted with class size[20]. The evidence suggests that the overall effect of external supervision is smaller in larger classes, in line with the second hypothesis.

Column (2) in the table shows that both the direct and the overall effect of external monitoring are higher in schools where the percentage of tenured teachers is higher. Typically, these are senior teachers with very secure jobs, who are less willing to adjust their teaching style to the needs of standardized tests and may therefore be more likely to engage in cheating and sabotaging.

Column (3) examines the interactions of the overall, direct and indirect effects with *ESCS*, the indicator of the average parental background in the class. If the incentives to engage in cheating were higher in classes with poor parental background, perhaps because teachers wish to altruistically compensate their students for their unfavourable initial conditions, we should find that the negative effect of external supervision is higher in these classes. Yet, there is no statistical evidence that this is the case[21].

Next, we ask whether the regional differences in the size of the effects of external monitoring are associated to the differences in the level of social capital[22]. Guiso, Sapienza and Zingales (2010) define social capital as civic capital, or as "...those persistent and shared beliefs and values that help a group overcome free rider outcomes..."(p.8). They report higher levels of social capital in Northern and Central Italy compared to the South.

We interact both the direct and the indirect effect of external monitoring with two measures of social capital at the provincial level taken from Guiso, Sapienza and Zingales (2004), the number of blood donations per 10,000 inhabitants in 1995 and the average electoral participation in the referenda held in Italy between 1946 and 1987. Since social capital is strongly correlated with local economic conditions, as shown in Figures 1.a-1.d, we also interact both effects with provincial GDP per capita and unemployment rates in 2009.

Results are shown in Table 9[23]. Column (1) in the table reports the estimates of the baseline model in the sub-sample of provinces for which data on social capital are available. These estimates are in line with those presented in Table 4. Column (2) and (4) show the interactions of the direct, indirect

[20] In this and in the following regressions the interacted variable is included also as an independent control.

[21] One possible explanation is that not only teachers, but also external examiners may be induced to engage in compensatory behaviour.

[22] In their seminal work, Putnam et al. (1993) links differences in the performance of local Italian governments to regional heterogeneity in social capital, measured in terms of local patterns of associationism, newspaper readership and political participation. Guiso, Sapienza and Zingales (2004) show that social capital is a key determinant of financial development, and Nannicini et al. (2012) study the impact of social capital on political accountability. Finally, Ichino and Maggi (2000) measure civicness in terms of shirking behaviour in the workplace and document large shirking differentials between Northern and Southern Italy.

[23] Descriptive statistics for these variables are shown in Table 2 – Panel B.

and overall effect of external monitoring with the two selected measures of social capital (blood donations and turnout at referenda, measured as deviations from sample means). We find that both the direct and the overall effect are smaller in schools located in provinces with a higher social capital. These qualitative results remain when we add to the regressions the interactions with provincial unemployment and GDP per capita (also measured as deviations from sample means, see columns (3) and (5)), although the effect of social capital is smaller.

Starting with Putnam's seminal contribution, several studies have suggested that Southern Italy has a lower endowment of "bridging" social capital, the form of social capital supportive of a more cohesive society and higher civicness[24], and is richer at the same time of "bonding" social capital, the type of social capital which reinforces family and clan ties in competition with the market and overall society and which is at the roots of the so called *amoral familism* (in the words of Banfield, 1958)[25]. We interpret the higher level of cheating observed in Southern Italy as the outcome of lower marginal costs of cheating due to lower "bridging" social capital, and/or of higher marginal benefits due to higher "bonding" social capital.

Conclusions

Test-based accountability systems in education may be gamed by students, teachers and school administrators in order to obtain higher measured levels of performance. This paper shows that having an external examiner who monitors test procedures has negative effects on the measured performance of tested classes and schools.These results are based on a natural experiment designed by the Italian national test administrator (INVALSI) to monitor test procedures in a random sample of Italian primary school classes. We have used random assignment to treatment to estimate both the direct and indirect effects of external monitoring. The former is based on the comparison of monitored and un-monitored classes within the same school and the latter on the comparison of un-monitored classes in schools with and without the external examiner.

The overall effect (direct plus indirect) of external monitoring is statistically significant and sizeable: depending on the grade, the presence of an external examiner reduces the percentage of correct answers in the class by 5.5 to 8.5 percent with respect to classes in schools with no external monitor. External monitoring spills over to un-monitored classes of the same school, but the size of this beneficial effect is rather small (about 20 percent of the overall effect).

Using additional supporting evidence on the psychological conditions of students before and during the test and on the distribution of answers within classes, we have concluded that the better performance of classes without the external examiner is due to the manipulation of test outcomes by

[24] Blood donations and referenda turnout measure bridging social capital.

[25] See Alesina and Ichino, 2009, for recent evidence.

teachers and/or students, and that the performance gap between monitored and un-monitored classes can be interpreted as a measure of the average intensity of cheating taking place in the latter.

While the direct negative effect of external supervision on test performance is not surprising, the presence of a small but statistical significant indirect negative effect is less expected. We have argued that this effect can be explained either by (irrational) fear of supervision or by a model where rational teachers administering the tests dislike excessive dispersion of test results within the school.

We believe that our results are useful for an economic assessment of external monitoring, which requires the evaluation of costs and benefits. To measure benefits, we need to ascertain whether external monitoring reduces cheating and by how much. Needless to say, using external examiners is not the only deterrence tool. Alternatives include re-shuffling the questions assigned to each students and computer – based tests. Reshuffling questions deters students from copying but does not strongly prevent cheating by teachers. Computer-based testing virtually eliminates cheating by teachers but it is quite costly, as it requires that each student is endowed with a computer. At the cost of 200 euro per workday, external examiners are rather cost-effective at reducing the manipulation of tests in a random sample of Italian schools. Yet, extending their use to the universe of tested schools seems complicated, not only because of the monetary costs involved but also because of the difficulty of finding enough qualified examiners.

References

Alesina, A. and Ichino, A., 2009. *L' Italia fatta in casa. Indagine sulla vera ricchezza degli italiani*. 1st ed. Milano: Mondadori.

Angelucci, M. and De Giorgi, G., 2009. Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption? *American Economic Review*, 99(1), pp. 486-508.

Banfield, E. C. (with L. Fasano), 1958. *The Moral Basis of a Backward Society*. 1st ed. Glencoe, IL: The Free Press.

Bokhari, F. A. S. and Schneider, H., 2011. School Accountability Laws and the Consumption of Psycho-stimulants. *Journal of Health Economics*, 30(2), pp. 355-372.

Cullen, J.B. and Reback, R., 2006. Tinkering Toward Accolades: School Gaming under a Performance Accountability System. In: Gronberg, T.J. and Jansen, D. W. (eds.), *Advances in Applied Microeconomics, 14*, pp.1-34.

Figlio, D. N., 2006. Testing, Crime and Punishment. *Journal of Public Economics,* 90(4), pp. 837-851.

Figlio, D. N. and Getzler, S.G, 2006. Accountability, Ability and Disability: Gaming the System. In: Gronberg, T.J. and Jansen, D. W. (eds.), *Advances in Applied Microeconomics, 14,* pp.35-49

Figlio, D. N. and Loeb, S., 2011. School Accountability. In: Hanushek, E. A., Machin, S. and Woessmann, L. (eds.), *Handbook of the Economics of Education*, 3, pp. 383-421.

Figlio, D. N., Winicki, J., 2005. Food for thought: the effects of school accountability plans on school nutrition, *Journal of Public Economics*, 89(2), pp. 381-394.

Guiso, L., Sapienza, P. and Zingales, L., 2004. The Role of Social Capital in Financial Development. *American Economic Review*, 94(3), pp. 526-556.

Guiso, L., Sapienza, P. and Zingales, L., 2010. Civic Capital as the Missing Link. NBER working Paper 15845.

Heckman, J.J., Lalonde, R. J. and Smith, J.A., 1999. The Economics and Econometrics of Active Labor Market Programs. In: Ashenfelter, O. C. and Card, D. (eds.), *Handbook of Labor Economics*, 3(1), pp. 1865-2097.

Hussain, I., 2012. Subjective Performance Evaluation in the Public Sector: Evidence from School Inspections. CEE Discussion Paper 135, London School of Economics.

Ichino, A. and Maggi, G. 2000. Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm". *Quarterly Journal of Economics*, 115(3), pp. 1057-1090.

INVALSI, 2010a. Sistema Nazionale di Valutazione – A.S. 2009/2010 - Rilevazione degli apprendimenti.

INVALSI, 2010b. Esami di Stato Primo Ciclo – A.S. 2009/2010 – Prova Nazionale. Prime Analisi.

Jacob, B. A., 2005. Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5), pp. 761-796.

Jacob, B. A. and Levitt, S., 2003. Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *Quarterly Journal of Economics*, 118(3), pp. 843-77.

Miguel, E. and Kremer, M., 2004. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica*, 72(1), pp.159-217.

Nannicini, T. et al., 2012. Social Capital and Political Accountability. *American Economic Journal: Economic Policy, forthcoming.*

Papke, L. E. and Wooldridge, J. M., 1996. Econometric Methods for Fractional Response Variables with an Application to 401(K) Plan Participation Rates. *Journal of Applied Econometrics*, 11(6), pp. 619-32.

Putnam, R. D. et al., 1993. *Making Democracy Work: Civic Traditions in Modern Italy*. 1st ed. Princeton, NJ: Princeton University Press.

Ricci, R. and Sestito, P., 2012. Il senso delle prove, *La Voce.Info*, July 6. 2012.

Tables and Figures

Table 1. Total and Sampled Number of Schools, Classes and Students. INVALSI SNV Test 2010

	Number of schools (total)	Number of classes (total)	Number of students (total)	Number of sampled schools	Number of sampled classes	Number of sampled students
Second Grade	7,700	30,175	555,347	1,385	2,000	39,299
Fifth Grade	7,700	30,476	565,064	1,385	2,000	39,643

Table 2. Mean and Standard Deviation of Covariates - Math Tests - V Graders

Panel A

	Mean	St Dev
Gender		
Missing (%)	0.01	0.10
Male (%)	0.50	0.50
Place of birth		
Missing (%)	0.04	0.20
Italy (%)	0.89	0.31
Citizenship		
Missing (%)	0.02	0.15
Italian (%)	0.89	0.32
First generation foreigner (%)	0.05	0.22
Second generation foreigner (%)	0.04	0.20
Pre-primary school		
Missing (%)	0.15	0.35
Yes (%)	0.83	0.37
Age		
Missing (%)	0.01	0.10
Older than regular (%)	0.03	0.16
Regular (%)	0.87	0.33
Younger than regular (%)	0.09	0.29
Math grade in previous semester (range:1-10)		
Missing (%)	0.07	0.26
1-4 (%)	0.00	0.04
5 (%)	0.04	0.20
6-7 (%)	0.38	0.48
8-10 (%)	0.51	0.50
Italian grade in previous semester (range:1-10)		
Missing (%)	0.07	0.25
1-4 (%)	0.00	0.04
5 (%)	0.04	0.19
6-7 (%)	0.41	0.49
8-10 (%)	0.48	0.50
Has own bedroom		
Missing (%)	0.03	0.17
Yes (%)	0.55	0.50
Has internet access		
Missing (%)	0.03	0.16
Yes (%)	0.76	0.43
Has an encyclopedia		
Missing (%)	0.03	0.16
Missing (%)	0.71	0.46
Has own desk		
Missing (%)	0.02	0.15
Yes (%)	0.85	0.36
Has a PC		
Missing (%)	0.03	0.16
Yes (%)	0.75	0.43
Has a place for homework		
Missing (%)	0.03	0.16
Yes (%)	0.84	0.37
Number of books at home		
Missing (%)	0.04	0.20
0-10 (%)	0.12	0.33
11-25 (%)	0.25	0.43
26-100 (%)	0.31	0.46
101-200 (%)	0.15	0.36
>200 (%)	0.12	0.33

	Mean	St Dev
Mother occupation		
Missing (%)	0.20	0.40
Unemployed or retired (%)	0.35	0.48
Employee (%)	0.31	0.46
Entrepreneur (%)	0.08	0.28
Middle manager (%)	0.06	0.23
Father occupation		
Missing (%)	0.22	0.41
Unemployed or retired (%)	0.04	0.19
Employee (%)	0.39	0.49
Entrepreneur (%)	0.25	0.43
Middle manager (%)	0.11	0.31
Mother education		
Missing (%)	0.21	0.41
Primary (%)	0.39	0.49
Secondary (%)	0.29	0.45
Tertiary (%)	0.11	0.32
Father education		
Missing (%)	0.22	0.42
Primary (%)	0.43	0.49
Secondary (%)	0.25	0.43
Tertiary (%)	0.10	0.30
Mother nationality		
Missing (%)	0.09	0.28
Italian (%)	0.80	0.40
Father nationality		
Missing(%)	0.09	0.29
Italian (%)	0.82	0.39
Private school	0.05	0.23
Full time schedule class	0.23	0.42
Number of students enrolled in class	19.00	4.65
Number of siblings		
Missing (%)	0.02	0.15
0 (%)	0.15	0.36
1 (%)	0.55	0.50
2 (%)	0.20	0.40
3 (%)	0.05	0.21
4 or more (%)	0.03	0.17
Lives with		
Missing (%)	0.02	0.15
Both parents (%)	0.86	0.35
One parent only (%)	0.06	0.24
Both parents alternatively(%)	0.05	0.22
Others (%)	0.01	0.08
Language spoken at home		
Missing (%)	0.04	0.21
Italian (%)	0.73	0.44
Dialect (%)	0.15	0.36
Other (%)	0.07	0.25
Help with homework		
Missing (%)	0.07	0.26
No homework (%)	0.01	0.07
No help needed (%)	0.20	0.40
Parents (%)	0.45	0.50
Siblings (%)	0.12	0.32
Private teacher (%)	0.03	0.17
Other (%)	0.04	0.20
No one (%)	0.09	0.28

Panel B (continued)

	Mean	St. Dev.
Tenured teachers in the school (%)	90.33	9.13
Class average ESCS index	-0.045	0.51
Class size	16.93	4.64
Blood donations	2.81	2.17
Average turnout at referenda (%)	80.28	8.37
Provincial unemployment rate (2009)	7.95	3.69
Provincial per capita GDP (2009)	23.84	5.60

Panel C

	Mean	St. Dev.
Math Test – V grade		
Score	0.65	0.19
Within-class standard deviation	0.14	0.04
Within-class coefficient of variation	0.23	0.09
Within-class bottom quartile	0.55	0.14
Within-class median	0.65	0.13
Within-class top quartile	0.75	0.12
Within-class Herfindal Index	0.53	0.15
Ranking based on Math scores	9.82	5.84
Ranking based on Math grades given by teachers in the previous semester	10.44	6.96
Anxiety Questions		
I was already anxious before starting the test	0.61	0.49
I was so nervous I couldn't find the answers	0.19	0.39
While answering , I felt like I was doing badly	0.50	0.50
While answering, I was calm	0.53	0.50
Absences from test (%)	0.11	0.10
Maths Test – II grade Score	0.62	0.20
Italian Test – V grade Score	0.70	0.18
Italian Test – II grade Score	0.65	0.23

Notes: The table reports the mean and standard deviation of the covariates included in the regressions (Panel A), the variables used in Section 7 (Panel B) and the dependent variables (Panel C). All numbers refer to the entire country. These statistics are based on individual, school and class level data. Except for the number of students enrolled in each class, the variables in Panel A have been categorized as dummy variables. Class size in Panel B refers to the number of students attending the test. Blood donations are the number of blood bags per 10,000 inhabitants in the province. Per capita GDP is measured in thousand euro. See the Appendix for further details.

Table 3 - Balancing Tests. First (between schools) and Second Stage (within schools) Randomization. - Math tests - V Graders.

Panel A

	Between schools	Within schools
Private school (%)	0.003	.
Full time schedule (%)	0.015	0.011
Number of students enrolled in class	0.079	0.425***
Gender		
Missing (%)	0.007***	0.020***
Male (%)	-0.005**	-0.004
Place of birth		
Missing (%)	-0.014***	-0.027***
Italy (%)	0.014***	0.027***
Citizenship		
Missing (%)	-0.008***	-0.013***
Italian (%)	0.008**	0.010***
First generation foreigner (%)	-0.001	0.000
Second generation foreigner (%)	0.001	0.002
Pre-primary school		
Missing (%)	-0.027***	-0.009*
Yes (%)	0.027***	0.010*
Age		
Missing (%)	0.007***	0.018***
Older than regular (%)	0.000	0.000
Regular (%)	-0.008***	-0.014***
Younger than regular (%)	0.002	-0.004**
Math grade in semester before the test		
Missing (%)	-0.021***	-0.009*
1-4 (%)	0.000	0.000*
5 (%)	0.001	0.000
6-7 (%)	0.010**	0.008*
8-10 (%)	0.011*	0.001
Italian grade in semester before the test		
Missing (%)	-0.021***	-0.008
1-4 (%)	0.000	0.000
5 (%)	0.000	0.001
6-7 (%)	0.006	0.003
8-10 (%)	0.014***	0.004

	Between schools	Within schools
Mother occupation		
Missing (%)	-0.014	-0.024***
Unemployed or retired (%)	0.008	0.012***
Employee (%)	0.003	0.004
Entrepreneur (%)	0.001	0.006**
Middle manager (%)	0.003	0.002
Father occupation		
Missing (%)	-0.014	-0.023***
Unemployed or retired (%)	0.001	0.001
Employee (%)	0.002	0.016***
Entrepreneur (%)	0.009*	0.005
Middle manager (%)	0.002	0.002
Mother education		
Missing (%)	-0.017	-0.028***
Primary (%)	0.008	0.019***
Secondary (%)	0.005	0.009**
Tertiary (%)	0.004	0.000
Father education		
Missing (%)	-0.018*	-0.025***
Primary (%)	0.013*	0.016***
Secondary (%)	0.001	0.008**
Tertiary (%)	0.003	0.001
Mother nationality		
Missing (%)	-0.018***	-0.014***
Italian (%)	0.015***	0.012**
Father nationality		
Missing (%)	-0.017***	-0.013***
Italian (%)	0.015***	0.009*

Panel B (continued)

	Between schools	Within schools
Has own bedroom		
Missing (%)	-0.006**	-0.009***
Yes (%)	0.000	0.004
Has internet access		
Missing (%)	-0.006**	-0.008***
Yes (%)	0.007**	0.008**
Has an encyclopedia		
Missing (%)	-0.006**	-0.008***
Yes (%)	0.005	0.016***
Has own desk		
Missing (%)	-0.005**	-0.008***
Yes (%)	0.005*	0.009***
Has a PC		
Missing (%)	-0.005**	-0.008***
Yes (%)	0.007**	0.011***
Has a place for homework		
Missing (%)	-0.006**	-0.008***
Yes (%)	0.006**	0.008**
Number of books at home		
Missing (%)	-0.007***	-0.008***
0-10 (%)	0.000	0.001
11-25 (%)	-0.004	-0.001
26-100 (%)	0.001	0.006*
101-200 (%)	0.004**	0.003
>200 (%)	0.006***	-0.001

	Between schools	Within schools
Number of siblings		
Missing (%)	-0.007***	-0.009***
0 (%)	-0.001	0.000
1 (%)	0.005*	0.008**
2 (%)	0.001	0.000
3 (%)	0.001	0.000
4 or more (%)	0.001	0.001
Lives with		
Missing (%)	-0.008***	-0.010***
Both parents (%)	0.008***	0.007**
One parent only (%)	-0.001	0.000
Both parents alternatively (%)	0.000	0.002
Others (%)	0.000	0.000
Language spoken at home		
Missing (%)	-0.008***	-0.009***
Italian (%)	0.004	0.007*
Dialect (%)	0.003	0.001
Other (%)	0.001	0.001
Help with homework		
Missing	-0.008***	-0.006**
No homework (%)	-0.001**	-0.001***
No help needed (%)	-0.001	0.005
Parents (%)	0.006*	0.001
Siblings (%)	0.003**	-0.002
Private teacher (%)	0.000	0.002
Other (%)	0.002	-0.001
No one (%)	-0.001	0.002

Notes: the table shows the point estimates of the balancing tests between and within schools. We compute school or class averages of individual variables and test for balancing using regressions (1) and (2). Full time schedule refers to schools offering this option in the between schools analysis and to the schedule of the single class in the within school analysis. While variables in Panel A are available for students in both grades, variables in Panel B are only available for fifth grade students. Standard errors for the second stage are adjusted for clustering at the school level. One, two and three stars for statistical significance at the 10, 5 and 1 percent level.

Table 4. The Effects of External Monitoring. Math Tests – V Grade. Dependent variable: Percentage of Correct Answers in the Class.

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect	-2.79***	-0.99***	-2.27***	-4.92***
	(0.25)	(0.28)	(0.48)	(0.50)
Indirect Effect	-0.81***	-0.70***	-0.73	-1.04*
	(0.28)	(0.27)	(0.45)	(0.61)
Overall Effect	-3.59***	-1.69***	-2.99***	-5.96***
	(0.29)	(0.31)	(0.54)	(0.60)
Observations	27,325	11,541	4,886	10,898
R-squared	0.15	0.2	0.15	0.14
Additional covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools	65.1	63.9	64.0	66.8

Notes: all regressions include the number of students enrolled in the class, regional dummies and regional dummies interacted with school size and with the number of fifth grade classes in the school. Additional covariates are shown in Table 2 - panel A. Estimates are weighted by class size. Standard errors adjusted for clustering at the school level in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table 5. The Effects of External Monitoring on Student Psychological Conditions. Math Tests – V Grade. Dependent variable: Percentage of Positive Answers in the Class.

	(1) I was already anxious before starting the test	(2) I was so nervous I couldn't find the answers	(3) While answering , I felt like I was doing badly	(4) While answering, I was calm
Direct Effect	0.25	-0.92***	-0.08	0.64
	(0.42)	(0.29)	(0.39)	(0.39)
Indirect Effect	0.25	0.01	0.36	-0.01
	(0.31)	(0.21)	(0.28)	(0.29)
Overall Effect	0.50	-0.90***	0.28	0.63*
	(0.41)	(0.28)	(0.38)	(0.38)
Observations	27,141	27,142	27,141	27,140
R-squared	0.07	0.11	0.1	0.07
Additional covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools	61.0	19.2	50.7	53.1

Notes: see Table 4. In each column, the dependent variable is the percentage of students in the class who agreed with the sentence reported at the top of the column. Students with missing answers have been dropped from the estimation sample (about 2 percent of the total). The estimates refer to the entire country.

Table 6. The Effects of External Monitoring on the Standard Deviation, the Coefficient of Variation and the Quartiles of the Distribution of Correct Answers within the Class. Math tests – V Grade.

	(1) Standard Deviation	(2) Coefficient of Variation	(3) Bottom quartile	(4) Median	(5) Top quartile
Direct Effect	0.76***	2.14***	-3.70***	-3.07***	-2.26***
	(0.09)	(0.21)	(0.31)	(0.29)	(0.27)
Indirect Effect	0.03	0.30	-0.55*	-0.56*	-0.61**
	(0.08)	(0.18)	(0.31)	(0.29)	(0.26)
Overall Effect	0.79***	2.44***	-4.26***	-3.63***	-2.88***
	(0.09)	(0.22)	(0.33)	(0.32)	(0.30)
Observations	27,325	27,325	27,325	27,325	27,325
R-squared	0.18	0.15	0.12	0.1	0.09
Additional covariates	Yes	Yes	Yes	Yes	Yes
Mean - Untreated Schools	14.1	22.8	55.6	65.6	75.2

Notes: see Table 4. The estimates refer to the entire country.

Table 7. The Effects of External Monitoring on the Heterogeneity of Answers in each Class. Math Tests – V Grade. Dependent Variable: Average Herfindhal Index in Each Class x 100.

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect	3.93***	1.24***	2.63***	7.32***
	(0.32)	(0.32)	(0.60)	(0.64)
Indirect Effect	0.82**	0.64**	0.51	1.22*
	(0.34)	(0.31)	(0.58)	(0.73)
Overall Effect	4.75***	1.88***	3.14***	8.54***
	(0.35)	(0.35)	(0.62)	(0.719)
Observations	27,325	11,541	4,886	10,898
R-squared	0.2	0.17	0.13	0.15
Additional covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools	52.8	57.3	55.7	46.9

Notes: see Table 4.

Table 8. Heterogeneous Effects of External Monitoring. Math Tests – V Grade. Dependent variable: Percentage of Correct Answers in the Class.

	(1) Interacted with Class Size	(2) Interacted with % Tenured Teachers	(3) Interacted with ESCS
Direct Effect	-3.41***	-1.34***	-2.65***
	(0.41)	(0.29)	(0.33)
Interacted Direct Effect	0.98*	-2.98***	-0.15
	(0.53)	(0.50)	(0.54)
Indirect Effect	-0.94***	-0.66**	-0.67**
	(0.36)	(0.29)	(0.31)
Interacted Indirect Effect	0.22	-0.19	-0.30
	(0.41)	(0.54)	(0.44)
Overall Effect	-4.35***	-2.00***	-3.32***
	(0.43)	(0.33)	(0.36)
Interacted Overall Effect	1.20**	-3.17***	-0.45
	(0.51)	(0.57)	(0.51)
Observations	27,325	26,313	27,323
R-squared	0.15	0.15	0.15
Additional covariates	Yes	Yes	Yes
Mean - Untreated Schools	65.1	64.9	65.1

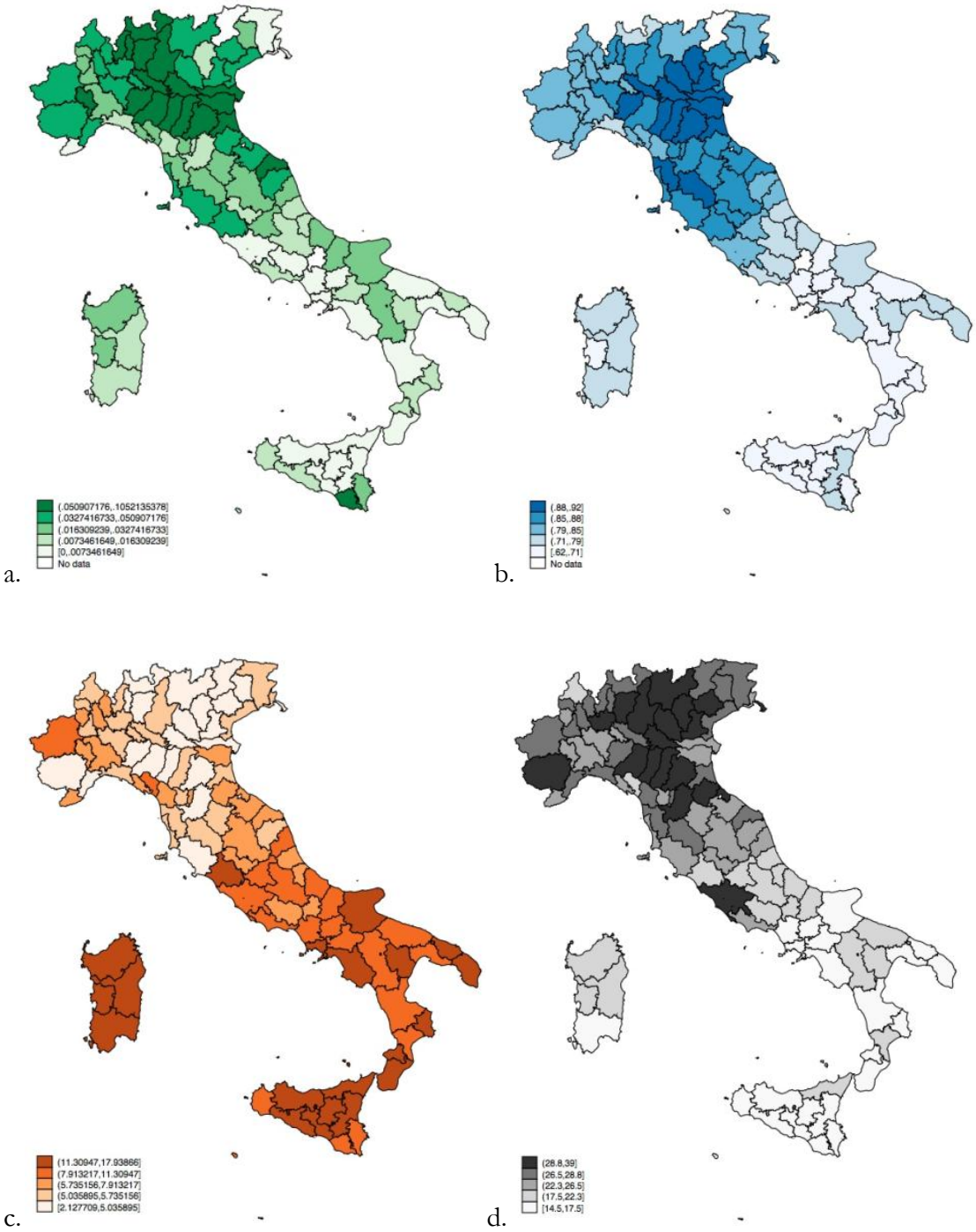
Notes: Interacted effects refer to the interactions between direct, indirect and overall effects and the variable listed at the top of each column. The interacting variable enters also as an independent covariate in the regression. Class size and the percentage of tenured teachers in the school are coded as dummy variables taking value one and zero when above and below the median. ESCS is coded as a dummy taking value one when below median and zero when above. The proportion of tenured teachers is not available for private schools (729 classes), for the public schools located in the Province of Trento (263 classes) and for five Sicilian public schools who did not transmit the information (20 classes). Average ESCS is not available for 2 classes. All regressions include the number of students enrolled in the class, regional dummies and regional dummies interacted with school size and with the number of fifth grade classes in the school. Estimates are weighted by class size. Standard errors adjusted for clustering at the school level in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table 9. Interacting External Monitoring with Measures of Social Capital. Math Tests – V Grade. Dependent variable: Percentage of Correct Answers in the Class

	(1)	(2)	(3)	(4)	(5)
	Baseline	Interacted with Blood Donations	Interacted with Blood Donations and Macro Variables	Interacted with Turnover at Referenda	Interacted with Turnover at Referenda and Macro Variables
Direct Effect	-2.78***	-2.48***	-2.64***	-2.63***	-2.69***
	(0.25)	(0.24)	(0.26)	(0.24)	(0.25)
Interacted Direct Effect		0.81***	0.41***	0.25***	0.14**
		(0.11)	(0.12)	(0.04)	(0.06)
Indirect Effect	-0.82***	-0.85***	-0.93***	-0.80***	-0.88***
	(0.28)	(0.26)	(0.29)	(0.26)	(0.20)
Interacted Indirect Effect		-0.06	-0.13	0.01	-0.02
		(0.12)	(0.13)	(0.04)	(0.07)
Overall Effect	-3.60***	-3.33***	-3.57***	-3.43***	-3.57**
	(0.30)	(0.28)	(0.31)	(0.28)	(0.30)
Interacted Overall Effect		0.75***	0.28**	0.26***	0.12*
		(0.13)	(0.14)	(0.04)	(0.07)
Observations	27,178	27,178	27,178	27,178	27,178
R-squared	0.15	0.15	0.15	0.15	0.15
Additional covariates	Yes	Yes	Yes	Yes	Yes
Mean - Untreated Schools	65.1	65.1	65.1	65.1	65.1

Notes: Interacted effects are the interactions between direct, indirect and overall effects and the variables listed at the top of each column. These variables enter as deviations from their sample means both in the interaction term and as an independent covariates in the regression. Social capital measures are not available for the provinces of Belluno and Isernia (147 classes). Macro variables: Per capita GDP and the unemployment rate in the province. All regressions include the number of students enrolled in the class, regional dummies and regional dummies interacted with school size and with the number of fifth grade classes in the school. Additional covariates are shown in Table 2 – panel A. Estimates are weighted by class size. Standard errors adjusted for clustering at the school level in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Figure 1. Geographical Distribution of Blood Donations, Average Turnout at Referenda, the Unemployment Rate and GDP per capita in the Italian Provinces.



Notes: Panel a): number of blood donations per 10,000 inhabitants in 1995. Panel b): average turnover at the referenda that took place between 1946 and 1989. Panel c): unemployment rate in 2009. Panel d) GDP per capita in 2009. The data are ordered by quintiles, with darker colours referring to the top quintile of the distribution.

Appendix

1) Tables

Table A.1. The Effects of External Monitoring. Italian Tests – V Grade. Dependent variable: Percentage of Correct Answers in the Class.

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect	-2.61***	-1.03***	-2.17***	-4.39***
	(0.20)	(0.21)	(0.42)	(0.39)
Indirect Effect	-0.67***	-0.38*	-0.81**	-0.99**
	(0.21)	(0.21)	(0.35)	(0.46)
Overall Effect	-3.28***	-1.41***	-2.98***	-5.37***
	(0.23)	(0.22)	(0.45)	(0.45)
Observations	27,369	11,557	4,894	10,918
R-squared	0.19	0.28	0.22	0.17
Additional covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools	70.0	70.2	70.1	69.7

Notes: see Table 4.

Table A.2. The Effects of External Monitoring. Math Tests – II Grade. Dependent variable: Percentage of Correct Answers in the Class.

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect	-4.20***	-1.57***	-3.09***	-7.50***
	(0.29)	(0.32)	(0.54)	(0.58)
Indirect Effect	-1.22***	-0.91***	-1.37**	-1.53**
	(0.33)	(0.34)	(0.60)	(0.74)
Overall Effect	-5.42***	-2.48***	-4.47***	-9.03***
	(0.34)	(0.36)	(0.58)	(0.69)
Observations	27,012	11,724	4,905	10,383
R-squared	0.11	0.08	0.09	0.08
Additional covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools	62.9	59.9	61.8	66.7

Notes: see Table 4.

Table A.3. The Effects of External Monitoring. Italian Tests – II Grade. Dependent variable: Percentage of Correct Answers in the Class.

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect x 100	-3.40***	-1.36***	-2.17***	-6.21***
	(0.28)	(0.34)	(0.51)	(0.54)
Indirect Effect x 100	-1.04***	-0.71**	-1.25**	-1.33**
	(0.28)	(0.31)	(0.53)	(0.62)
Overall Effect x 100	-4.44***	-2.07***	-3.42	-7.54***
	(0.29)	(0.34)	(0.56)	(0.58)
Observations	27,025	11,721	4,911	10,393
R-squared	0.13	0.2	0.16	0.11
Additional covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools	65.9	65.0	66.2	66.7

Notes: see Table 4.

Table A.4. GLM estimates of the Effects of External Monitoring. Math Tests – V Grade. Dependent variable: Percentage of Correct Answers in the Class.

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect	-2.74***	-0.97***	-2.25***	-4.73***
	(0.25)	(0.28)	(0.47)	(0.48)
Indirect Effect	-0.80***	-0.70***	-0.72	-1.04*
	(0.28)	(0.27)	(0.45)	(0.60)
Overall Effect	-3.54***	-1.67***	-2.97***	-5.77***
	(0.29)	(0.30)	(0.53)	(0.57)
Observations	27,325	11,541	4,886	10,898
Additional covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools	65.1	63.9	64.0	66.8

Notes: see Table 4.

Table A.5. The Effects of External Monitoring. Math Tests – V Grade. Dependent variable: Percentage of Correct Answers in the Class. Finite Population Correction.

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect	-2.89***	-1.08***	-2.35***	-5.05***
	(0.12)	(0.14)	(0.23)	(0.24)
Indirect Effect	-0.83***	-0.71***	-0.70***	-1.06***
	(0.13)	(0.13)	(0.21)	(0.27)
Overall Effect	-3.72***	-1.79***	-3.05***	-6.11***
	(0.14)	(0.14)	(0.25)	(0.28)
Observations	27,325	11,541	4,886	10,898
R-squared	0.15	0.19	0.15	0.15
Additional covariates	Yes	Yes	Yes	Yes
Mean - Untreated Schools	65.1	63.9	64.0	66.8

Notes: see Table 4.

Table A.6. The Effects of External Monitoring. Math Tests – V Grade. Dependent variable: Percentage of Correct Answers in the Class. Without Covariates.

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect	-2.82***	-0.85***	-2.04***	-5.29***
	(0.26)	(0.30)	(0.49)	(0.52)
Indirect Effect	-0.70**	-0.82***	-0.46	-0.70
	(0.30)	(0.31)	(0.51)	(0.65)
Overall Effect	-3.52***	-1.68***	-2.50***	-5.99***
	(0.31)	(0.34)	(0.58)	(0.64)
Observations	27,325	11,541	4,886	10,898
R-squared	0.03	0.01	0.01	0.03
Additional covariates	No	No	No	No
Mean - Untreated Schools	65.1	63.9	64.0	66.8

Notes: see Table 4. Each regression includes the number of students enrolled in the class, regional dummies and regional dummies interacted with school size and with the number of fifth grade classes in the school.

Table A.7. The Effects of External Monitoring. Math Tests – V grade. Dependent variable: Percentage Absent from the Test

	(1) Italy	(2) North	(3) Centre	(4) South
Direct Effect	-0.53**	-0.50	-0.47	-0.55
	(0.24)	(0.40)	(0.47)	(0.40)
Indirect Effect	-0.10	0.44	-0.44	-0.51
	(0.24)	(0.36)	(0.42)	(0.44)
Overall Effect	-0.63**	-0.06	-0.91**	-1.06**
	(0.25)	(0.40)	(0.46)	(0.42)
Observations	27,325	11,541	4,886	10,898
R-squared	0.03	0.02	0.03	0.03
Additional covariates	No	No	No	No
Mean - Untreated Schools	11.0	10.4	11.7	11.4

Notes: see Table 4. The only covariates still included in the models are the number of students enrolled in the class, regional dummies and regional dummies interacted with school size and with the number of fifth grade classes in the school.

2) External examiners.

External examiners are selected by the regional education offices using criteria defined at the national level, from a pool of potential candidates composed by teachers and school principals, most of them retired. Eligible candidates must have personal characteristics that facilitate a fair collaboration with the school principal and the teachers involved in the test, should have a good knowledge of the evaluation procedure and should be familiar with the software and the procedure to transmit data to INVALSI.

Eligibility requires that examiners did not work during the two years before the test in the same municipality or in the same school they are going to supervise. If they are still active as teachers, they must be employed in a non-primary school. INVALSI conducted some investigation about possible cases of collusion between external examiners and school principals or teachers and did not find evidence of misconduct. Once appointed, external examiners need to coordinate with the school principal to prepare for the test. External examiners generally worked for two days and earned 200 euro per working day.

3) *Sampling procedure.*

The sampling procedure is a two-stage design and was taken from the IEA TIMMS survey, which INVALSI manages for Italy. Sampling takes place separately in each region. In the first stage, a pre-specified number of schools was randomly drawn from the population of schools located in the region. Schools with less than 10 students were excluded from the population and the rest were listed in a spreadsheet with the corresponding number of enrolled students in the second and fifth grades, which is the relevant measure of school size. The sampling method adopted is a PPS – probability proportional to size: the probability that each school is randomly sampled is proportional to school size. Practically, a software randomly samples schools from the sampling frame.[26] Only 5 schools have been replaced from the original sample. This low replacement rate is due to the fact that participation and compliance with INVALSI procedures are compulsory because of the law. The second stage of the sampling procedure is a simple random sampling of classes within the sampled schools. One or two classes per grade, depending uniquely on school size, were randomly selected from each sampled school. No negotiation between school principals and INVALSI occurred to determine the selected classes.

The PPS technique implies that larger schools have a higher probability of being sampled than smaller schools. However, this difference in the selection probabilities is largely offset at the second

[26] Additional details on the sampling of schools can be found at the IEA TIMSS and PIRLS 2011 webpage http://timssandpirls.bc.edu/methods/pdf/TP_Sampling_Design.pdf

stage of sampling by selecting a fixed number of classes with equal probability from the sampled school. Classes in large schools with many classes in the target grade have a lower probability of selection than classes in smaller schools that have just one or two classes.

4) From the initial dataset to the final sample

Our data are drawn from the 2010 wave of the INVALSI SNV survey of educational achievements in Italian primary schools. These data are freely available from INVALSI. In this section of the Appendix we briefly describe our handling of the data.

1) We exclude Valle d'Aosta and the Province of Bolzano, because all classes in these areas were assigned to external monitoring.
2) We drop schools where there is a different number of second and fifth grade classes assigned to monitoring, because this outcome is inconsistent with the sampling scheme.
3) We drop classes with less than five students and schools with a single class per grade or with two classes if both were assigned to monitoring.

To illustrate the effects of these actions, we consider the Math test for fifth graders. For this group, the population consists of 7,700 schools, 30,476 classes and 565,064 students. Our initial dataset includes 7,541 schools, 29,811 classes and 491,421 non-disabled students in schools with more than ten students (smaller schools are excluded from testing) who were present during the testing day. Dropping data for the provinces of Aosta and Bolzano reduces the total number of schools to 7,502, with 29,647 classes and 489,396 students. Elimination of treated schools where there is a different number of second and fifth grade classes leaves us with 489,126 students allocated in 29,629 classes of 7,498 schools. Purging out classes with less than 5 students leaves us with 28,677 classes in 7,452 schools and a total of 486,531 students. After dropping schools with a single class in the grade or with two classes if both are treated we obtain our estimation sample, which consists of 6,108 schools, 27,325 classes and 462,570 students.

5)Other data

Unemployment and per capita GDP data refer to year 2009 and are drawn from EUROSTAT regional statistics database. Data on blood donations and the average turnout at referenda are from Guiso, Sapienza and Zingales (2004). The original data have been re-classified to match INVALSI classification, which includes 103 provinces

Chapter 2

Selection and the Age - Productivity Profile. Evidence from Chess Players*

by

Marco Bertoni
(University of Padova and CEP, LSE)

Giorgio Brunello
(University of Padova, IZA and CEsifo)

Lorenzo Rocco
(University of Padova)

Abstract
We use data on professional chess tournaments to study how endogenous selection affects the relationship between age and mental productivity in a brain-intensive profession. We show that less talented players are more likely to drop out, and that the age-productivity gradient is heterogeneous by ability, making fixed effects estimators inconsistent. We correct for selection using an imputation procedure that repopulates the sample by applying to older cohorts the self-selection patterns observed in younger cohorts. We estimate the age-productivity profile on the repopulated sample using median regressions, and find that median productivity increases by close to 5 percent from initial age (15) to peak age (21), and declines substantially after the peak. At age 50, it is about 10 percent lower than at age 15. We compare profiles in the unadjusted and in the repopulated sample and show that failure to adequately address endogenous selection in the former leads to substantially over-estimating productivity at any age relative to initial age.

Keywords: aging, productivity, mental ability.
JEL codes: D83, J14, J24.

*The authors are grateful to an Editor, three anonymous referees, Guy Michaels, Mario Padula, RoopeUusitalo, Jan Van Ours and to the audiences at seminars in London (LSE), Padova, Turin (EALE) and Venice for comments and suggestions. We also thank Michele Bertoni for technical help with data collection. Financial support from the University of Padova (Grant CPDA093857 – PercorsiLavorativi e InvecchiamentoAttivo) is gratefully acknowledged. All errors are our own.

Introduction

There is a broad perception that mental ability declines with age, and not just for humans.[1] Unless experience, knowledge, motivation and effort can fully compensate the decline in ability, productivity is also bound to decline. In many developed countries, population is ageing. If individual productivity declines with age, overall productivity will also decline, with important macroeconomic implications.

In spite of the important implications for modern economies, surprisingly little is known about the relationship between age and productivity, and the little we know is not pointing unambiguously in the same direction. Skirbekk, 2003, reviews the empirical literature and concludes that productivity follows an inverted U-shaped profile, with significant decreases taking place from around age 50.[2] Van Ours, 2009, on the other hand, finds that while physical productivity does decline after age 40, mental productivity – measured by publishing in economics journals – does not decline even after age 50. Finally, Borsch-Supan and Weiss, 2007, use data on production workers of a large German car manufacturer and conclude that productivity does not decline at least up to age 60.[3]

Measuring the effects of age on productivity is difficult. First, it is hard to find reliable measures of individual productivity. Second, in many jobs individual productivity should include also the effects on the productivity of others, either because of knowledge spill-overs or because some jobs involve a relevant team component. Third, the relationship between age and productivity in observed samples is often affected by endogenous selection. If more productive workers are more likely to stay in their jobs, for instance because they retire later (see Myck, 2010), selection may induce a spurious positive correlation between age and productivity.

In this paper, we investigate the effects of endogenous selection on the age-productivity profile by using data on professional chess players. Focusing on chess players has important advantages. First, we can compute a quality - adjusted measure of individual productivity by looking at wins and draws in professional tournaments, weighting each result with the measured strength of the opponent. Second, chess is a purely individual activity, differently from most professional activities where team work and spill-overs among agents influence individual output. Because of this, our measure of productivity is accurate.

Using longitudinal data on professional chess players with a rating provided by FIDE, the international chess federation, and on all the official FIDE tournaments played worldwide between 2008 and 2011, we show that participation to these tournaments is characterized by substantial attrition, and that the relationship between age and productivity is heterogeneous by ability. When

[1]See The Economist, 2004 and Bloom and Sousa-Poza, 2013.

[2]Recent contributions in this area that use individual productivity data include Weinberg and Galenson, 2005, and Castellucci, Padula and Pica, 2010.

[3]Pekkarinen and Uusitalo, 2012, look at the population of Finnish blue-collar employees and use piece-rate wages as proxies for output.Their findings confirm that labour productivity stays roughly constant after age 40.

productivity is not separable in terms of age and ability, and there is selective attrition, commonly used fixed effects methods fail to deliver consistent estimates of the age-productivity profile.[4]Following Olivetti and Petrongolo, 2008, we address the problems associated to endogenous attrition with imputation, and reconstruct the population of chess players by applying to older cohorts the self-selection patterns observed in younger cohorts. Conditional on an assumption of stationarity in the selection process across cohorts, we obtain that median productivity by age in the repopulated data is equal to productivity in the absence of endogenous attrition.

We compare median age-productivity profiles in the unadjusted and in the repopulated sample, after netting out country and cohort effects, and find that the differences are stark. In the former case, productivity peaks at age 24 and declines almost monotonously thereinafter. At peak age, it is 18percent higher than at baseline age (15). At age 50, it is about at the same level as the baseline. In the latter case, the peak occurs earlier, at age 21, and the increase from the baseline to the peak is much smaller (about 5 percent) than in the unadjusted sample. Furthermore, median productivity at age 50 is only about 90 percent of baseline productivity. These results point out that failure to account for endogenous selection and for the fact that less talented players tend to leave the game can lead both to substantially over-estimating productivity at all ages relative to baseline age and to obtain distorted estimates of the age-productivity profile.

Several studies (see Skirbekk, 2003) have shown that the decline of mental abilities from early adulthood is a universal phenomenon. Unless the acquisition of skills and experience on the job outweighs this decline, productivity in cognitive tasks is also likely to fall with age. In this study, we find that the median productivity of professional chess players is significantly lower at age 40 than at age 20. This evidence from professional chess, a brain-intensive activity, suggests that better skills and longer experience cannot offset the decline in numerical and reasoning abilities.

We are aware that chess is a rather special cognitive profession and that results cannot be easily generalized to economically more important tasks. We believe, however, that the contribution of our paper is partly methodological and goes beyond the specific environment characterizing professional chess: we show that productivity at any age is substantially over-estimated and that the estimates of age-productivity profiles are distorted when endogenous selection and heterogeneity by ability are not properly addressed.

The paper is organized as follows. In Section 1 we introduce our measure of productivity for chess players. Section 2 presents the data. The estimation strategy and the imputation method are discussed in Sections3 and 4. Results are in Section 5. Conclusions follow.

[4]SeeGobel and Zwick, 2011, for a discussion of estimation methods in this area of research.

1. Measures of Ability and Productivity for Professional Chess Players

Ranking players has been a critical issue in chess until the 1960s, when the ELO rating system was introduced by FIDE, the International Chess Federation. This system was developed by the Hungarian mathematician Arpad Elo and is based on a Thurstonian model for paired comparisons (see Thurstone, 1927). In this section, we argue that ELO is not a measure of individual productivity but rather an indicator of individual (relative) ability in the game of chess.

In the ELO system, the latent ability of player i, μ_i,is assumed to be normally distributed with mean s_i and standard deviation arbitrarily set at 200.[5] Let the outcome of a match between players *i* and *j* be the random variable $z_{ij}=\mu_i-\mu_j$. Player *i* wins if $z_{ij} > 0$. With independent abilities, the probability of winning is $p_{ij} = \Phi\left(\frac{s_i - s_j}{\sigma}\right)$, where Φ is the cumulative distribution function of a standard normal random variable and $\sigma=200\sqrt{2}$.[6]

The expected ability s_i of player *i* is estimated using the outcomes of the games she plays. Players are initially classified as unrated.[7] Starting from their first official ELO score, s_{i0}, the score after game *g* is given by $s_{ig+1} = s_{ig} + K(w_{ij} - p_{ij})$, where w_{ij} is equal to 1if player *i* wins, to 0.5 if she draws and to 0if she loses the match, p_{ij} is the expected winning probability of player *i* against player *j*, and *K* is a scale factor which weights the importance of a single game with respect to her entire previous career. This weight declines with the number of games played and with the ELO score.[8]

The updating rule adjusts the ELO score when actual performance in the game differs from expected performance. When the current ELO perfectly predicts average actual performance, no further update occurs. Since only unexpected wins and losses matter in the updating mechanism, ELO cannot be considered a measure of productivity at chess, which depends on realized rather than unexpected wins and draws. To illustrate, a player can be very productive in terms of having a high winning rate and yet experience no change in ELO if these wins are expected.[9]To us, rather

---

[5]The normality assumption is based on observational data collected by Arpad Elo on the distribution of individual chess performance (see Gransmark and Gërdes, 2010).

[6] For example, consider two players with s_i-s_j=200. In this case, the likelihoods that players *i* and *j* win are equal to $\Phi(200/200\sqrt{2})=0.76$ and 0.24 respectively.

[7] The results of their first games and the ELO score of their opponents determine a provisional rating. The following conditions are required to obtain a rating (see FIDE, 2012): 1) having played in at least one official FIDE tournament; 2) having completed a minimum of nine games against rated players and having scored at least one point against them (i.e., having won a match or having drawn two);3) the initial score ought to be above a minimum rating floor, equivalent to 1400 ELO points for players in our sample, who obtained their first rating before 2009.

[8]In practice, K = 30 for a player who has completed less than 30 games, K = 15 for players with a score lower than 2400 and K = 10 once the rating reaches 2400 and she has completed at least 30 games (see Glickman, 1995, for details). Using the example in footnote 2, if player *i* wins, her ELO score increases by 0.24*K, while if she loses her ELO decreases by -0.76*K.

[9]Furthermore, two players with the same initial ELO but different K factors (i.e. different experience) have different ELO adjustments even if their game results are the same, making the use of ELO as a measure of productivity even more problematic.

than a measure of productivity, ELO is a measure of relative ability at chess at a given point in time: it predicts ex-ante how likely a player is to win when he plays against another one, but it does not measure winning intensity.[10]

We therefore distinguish between ELO and productivity, *Y*: in our view, the former is an estimate of relative ability, and the latter is measured as the weighted sum of wins and draws divided by the number of played games G_{it}

$$Y_{it} = \frac{\sum_{j=1}^{G_{it}} [I(win_{ij}) * ELO_j + \frac{1}{2} I(draw_{ij}) * ELO_j]}{G_{it}} \qquad [1]$$

where *i* is for the player and *t* for the year, and $I(win_{ij})$ and $I(draw_{ij})$ are dummies equal to 1 when either a win or a draw occur. Each win has weight equal to 1, each draw is weighted 0.5 and each loss has zero weight. This measure of productivity is quality adjusted because each win or draw is weighted with the relative quality of the opponent, measured by the ELO score. Since the weighted sum of wins and losses is divided by the number of games played, Y_{it} is the productivity per match.[11]

Productivity as defined in (1) is used in our baseline regressions. We also experiment, however, with two alternative definitions. In the former, we replace the ELO score of the opponent in the tournament with the average ELO score of all opponents met during the year. In the latter, we add up wins and draws without using weights.

2. The Data

Our data consist of a main and an auxiliary sample. The main sample has information on all official FIDE tournaments played worldwide between 2008 and 2011, that we have downloaded from the FIDE online archive.[12]Each tournament record reports the results of all the games played by each participant (wins, losses or draws), and his ELO score before and after the tournament.

Our initial sample consists of male players with a FIDE rating, for whom we know the ELO score and the national federation, who were born between 1948 and 1993, were listed in FIDE by 2008 and have played in at least one FIDE tournament between 2008 and 2011.From this sample, we drop "casual" players, who obtained an official rating for the first time in 2008 and have played only

[10] ELO changes faster at younger ages, because the updating mechanism generates larger variations when the initial ELO is lower and because younger players try to fill their ability gap with more experienced players by learning, training and accumulating experience in tournaments. Hence, it is an informative but imperfect measure of innate talent at chess.

[11]Our weighting system implies that playing two games against players of a given strength and winning both is equivalent to playing two games against opponents twice as strong and winning only one game. It also implies that winning one game against a player with ELO score *x* yields more in terms of productivity than drawing one game against a player with ELO equal to 2*(x-ε).

[12] As of December 2012, the web address of this archive is http://ratings.fide.com.

in 2008. For the remaining players, we only consider the outcomes of games played against rated players, both because we do not have a measure of ability for unrated opponents and because games against these opponents do not count for rating.[13] We also drop players belonging to national federations with less than 30 affiliates. Our final sample consists of 40,545 players aged between 15 and 60, who were listed in 2008 and remained in the FIDE list from a minimum of 1 to a maximum of 4 years.[14]Since our panel is not balanced, we end up with 140,074 observations.

Table 1 presents descriptive statistics on age, the number of games played and productivity. Average age is 38.09,the annual average number of games played by active players is 17.45 (range from 1 to 289), and average productivity is equal 972.39, with a range between 0 and 2551.[15]Figure 1 shows the distribution of annual productivity, which exhibits a peak at zero (3.6% of observations), due to players who have never won or drawn a game in a single year, and an upper tail with few players having very high productivity. Table 1 also includes the means of several variables at the federation-by-year level, that will be discussed in Section3: the number of tournaments organized in the year, the GDP per capita in real PPP 2005 (thousand) dollars, the number of internet users per 100 inhabitants, the rate of growth of GDP per capita when individuals are aged 15 to 25, and the average number of internet users per 100 inhabitants for the same age interval.[16]Except for the number of tournaments, which are derived from our data, the other variables are drawn from the World Bank World Development Indicators, the Maddison Project Database and the Statistics of the International Telecommunication Union.

We plot the empirical relationship between age and median productivity in Figure 2 (the continuous line), after netting out country fixed effects and country-specific cohort and period effects.[17]We find that median productivity increases by 18 percent, from 925.3 at age 15 to 1092.0 at age 24, and then declines almost linearly until 869.2 at age 60, when it is 6 percent below the level at age 15. In the figure, we also plot median productivity for the two sub-groups of players with an ELO score in 2006 above and below 2000 (dashed lines)[18]. We notice two things. First, the two dashed lines are not parallel, as productivity declines more steeply amongst the less talented (bottom line). This

[13] In the few cases where annual productivity is missing in either 2009 or 2010 but not in 2008 and 2011, we estimate missing values by interpolation.

[14]The number of players enrolled in the lists in 2008 is 40,545. Of these, 37,396 are still present in 2009, 33,475 in 2010 and 28,658 in 2011.

[15]The relatively large value of average productivity is justified by our weighting wins and draws with the ELO of the opponent. In our sample, average ELO is 2073.7.

[16]In practice, we match to each individual the average rate of GDP growth and the average number of internet users in the period when he is (was) 15 to 25.

[17] For the sake of comparability with the rest of the paper, Figure 2 is obtained by grouping players' age across three adjacent years, starting from age 15. Thus players aged between 15 and 17 are assigned aged 15, players aged between age 18 and 20 are assigned age 18 and so on. A Figure based on ungrouped age is available upon the authors, but results are qualitatively the same.

[18]According to the US Chess Federation, a score equal to 2000 separates experts from "regular" chess players.We consider ELO in 2006, rather than at the beginning of our short panel starting in 2008, to have a predetermined measure of ability, not affected by the performance of players between 2008 and 2011. The data are from the FIDE lists. For young players aged 15 or 16 in 2008 we use the first available ELO score.

implies that the relationship between age and productivity is heterogeneous by ability. Second, median productivity in the full sample raises sharply in the early stages, partly because less talented players, who are also less productive, are more likely to leave the sample as time goes by, thus altering the composition of the pool of players by ability at different ages.

The auxiliary sample is a longitudinal dataset that tracks until 2011 all rated players that were included in the FIDE lists in 2001, and for each player has information on the federation, the year of birth, gender and the ELO score. Productivity, however, is missing. We use this dataset to better document the endogenous selection of professional players, who enter and exit the FIDE lists every year. We define as "stayers" the rated players still active between 2009 and 2011, and as "dropouts" those not active between 2009 and 2011. Setting at 100 the number of players listed in 2001, we find that dropouts at the end of the window of observation were25.78, about a quarter of the entire pool. We define the dummy DR (dropout) as equal to one if the player is a dropout and to zero otherwise, and regress this dummy on country effects, age and the ELO score in 2001, using a Probit model. We find that the probability of dropping out declines with age and is higher among the players with a lower initial ELO score. In particular, we estimate that adding 100 points to the ELO score is associated to a 6.8 percentage points reduction in the dropout rate, a substantial effect.[19]

3. The Empirical Strategy

Age can affect productivity by influencing mental ability, skill accumulation and experience, learning, motivation and effort. Outside options available to chess players may also change with age and affect the decision to stay on as a professional player or to drop out of FIDE lists.[20]Because of all these reasons, productivity Y is a function of age A and mental ability μ. As documented in the previous section and tested at the end of this section, this function is not separable in terms of age and ability. Therefore, we write

$$Y_{it} = Y(A_{it}, \mu_{it}) = \pi_0 + \pi_1 A_{it} + \pi_2 \mu_{it} + \pi_3 A_{it} \mu_{it} \qquad [2]$$

where $\pi_3$is positive if individuals with higher ability are better capable of accumulating skills as they age.

Mental ability μ_{it}consists of time invariant innate talent α and a component that declines with age

[19] Detailed results are available from the authors upon request.

[20] As remarked by a chess-expert referee, aging in chess is not one process, but many processes, some of which run in opposite directions. This is the case of evaluation and calculation. Calculation in chess is the I-do-this-then-he-does-that-then-I-do-this part, which is very strenuous. Evaluation is the ability to judge who is favoured in a certain position. This is done partly by analogy with known positions, and more experienced players have a larger mental "database" of positions to draw from.

$$\mu_{it} = \alpha_i - \rho A_{it} \qquad [3]$$

Using [3] into [2], we can express productivity as a function of age, innate talent, and the interaction of talent with age. We describe the empirical relationship between productivity, age, talent and other covariates as follows

$$Y_{it} = \beta_0 + \sum_{d=0}^{D} \beta_d A_{it}^d + \beta_x X_{it} + \gamma \alpha_i + \delta \alpha_i A_{it} + \varepsilon_{it} \qquad [4]$$

where $\sum_{d=1}^{D} \beta_d A_{it}^{\ d}$ is a polynomial of order d in age, ε is a random error, innate talent α has unconditional zero mean and is orthogonal to age in the population, and X_{it} is a vector of country – specific period and cohort effects.

Since age, period and cohort are linearly dependent and cannot be simultaneously controlled for non-parametrically, we choose to capture period and cohort effects with period-by-country and cohort-by-country variables. The former include the real GDP per capita, the share of internet users per 100 inhabitants and the number of tournaments organized in the year by each federation. GDP per capita is expected to capture access to resources and to affect the outside option of chess players; the share of internet connections captures access to the internet and to the training opportunities offered by the web; finally, the number of tournaments organized by the each federation is a measure of the supply of opportunities to play. The latter include the average rate of growth of GDP and of internet usage when each player was aged between 15 to 25.

The orthogonality of talent and age implies that the conditional mean of [4] in the population is given by

$$E\left[Y_{it} \mid A_{it}, X_{it}\right] = \beta_0 + \sum_{d=0}^{D} \beta_d A_{it}^d + \beta_x X_{it} \qquad [5]$$

If we had population data, we could estimate the relationship between age and mean productivity by ordinary least squares, and the interaction between age and ability could safely be omitted. The conditional mean talent in the population and in the observed sample do not coincide, however, when individual players select in and out of the sample in a non-random way and returns are heterogeneous by ability. In the case of professional chess players, the decision to stay or leave the FIDE lists depends both on individual talent and on age, as discussed in Section 2.

Therefore, in the unadjusted sample the conditional mean $E\left[\alpha_i \mid A_{it}, X_{it}\right]$ is different from zero and we have that

$$E\left[Y_{it} \mid A_{it}, X_{it}\right] = \beta_0 + \sum_{d=0}^{D} \beta_d A_{it}^d + \beta_x X_{it} + E\left[\alpha_i \mid A_{it}, X_{it}\right]\left(\gamma + \delta A_{it}\right) \qquad [6]$$

The conditional expectation of productivity depends both on the (nonzero) conditional mean of innate talent and on the interaction of this mean with age. When $\delta \neq 0$, productivity returns to age are heterogeneous by talent. In this case, applying a fixed effects estimator to the raw data – as done by Castellucci, Padula and Pica, 2010 – does not eliminate the bias due to selection because the within-transformation only removes the linear component of talent. To fully remove the selection bias, one needs to apply the within-player transformation to first-differenced data, as done for instance by Pischke, 2001, in his paper on the returns to training in Germany. This transformation, however, has the drawback that the linear term of the age polynomial in [4] cannot be identified when $E\left[\alpha_i \mid A_{it}, X_{it}\right]$ is different from zero. We therefore turn to a method based on imputation, which allows us to reconstruct the population of chess players that would have been observed in the absence of self-selection.

Before considering this method, however, we conclude this section by introducing a Hausman test of the hypothesis that the age-productivity gradient in our data is not heterogeneous by ability. If this is the case, $\delta = 0$ in equation [4] (the null hypothesis). Under the null, first differences of [4] yield $\Delta Y_{it} = \sum_{d=1}^{D} \beta_d \Delta A_{it}^{\ d} + \beta_x \Delta X_{it} + \Delta \varepsilon_{it}$, which can be consistently estimated on first-differenced data by ordinary least squares or by fixed effects models, with the former being more efficient than the latter. When the null does not hold, first differences of [4] yield $\Delta Y_{it} = \sum_{d=1}^{D} \beta_d \Delta A_{it}^{\ d} + \beta_x \Delta X_{it} + \delta \alpha_i + \Delta \varepsilon_{it}$, which can only be consistently estimated by fixed effects. We follow Imbens and Wooldridge, 2007, and perform the Hausman test in a setup where standard errors are clustered with respect to the individual, by augmenting the set of regressors in $\Delta Y_{it} = \sum_{d=1}^{D} \beta_d \Delta A_{it}^{\ d} + \beta_x \Delta X_{it} + \Delta \varepsilon_{it}$ with the deviations of each explanatory variable from its individual mean. We then test whether these additional regressors are jointly significant. Since we find that the p-value of the joint F test is equal to 0.0002,we reject the null hypothesis ($\delta = 0$) and conclude that returns are heterogeneous by talent.

4. Imputation

Our imputation method assumes that the self-selection process affecting the relationship between age and productivity depends on age and ability, but not on the cohort of birth. By virtue of this

assumption, we are able to reconstruct the original population of older cohorts by using the patterns of self-selection observed in younger cohorts. After completing imputation, we estimate the relationship between age and productivity on the repopulated sample using median regressions. As remarked by Olivetti and Petrongolo, 2008, the attractive feature of median regressions for imputation methods is that estimates only depend on the position of individuals with respect to the median observation, and not on the specific imputed value of the outcome variable. Thus, as long as observations with missing productivity are correctly imputed to the side of the median where they belong, median regressions retrieve the true parameters of interest.[21]

Since the size of our sample in the main dataset is relatively small when we consider age-by-period cells, we pool individuals born in three contiguous years of birth into a single cohort. Therefore, using 2008 as our reference, our youngest cohort - c_{15} - is aged 15 to 17 and our oldest cohort - c_{60} - is aged 60 to 62. As each cohort consists of three contiguous years of birth, we only retain the initial and final year in our sample (2008 and 2011), that we denote *t=0* and *t=3* respectively, to avoid overlaps between cohorts.

Our imputation strategy requires the following three assumptions:

Assumption A0 *(normality)*. For each cohort, the distribution of innate talent in the population is normal and has zero mean.

Assumption A1 (*anchoring*). Cohort c_{15} at time *t=0* is not self-selected. Therefore, $E[\alpha_i|c_{15}, t=0] = Med(\alpha_i|c_{15}, t=0) = 0$.[22]

Assumption A2 (*stationarity*).Median productivity at a given age is invariant across cohorts. Let productivity net of the effect of the exogenous controls included in vector X_{it} be $\bar{y}_{ict}$, where *c* is the cohort. Then $Med(\bar{y}_{itc} \mid c=k, t=3) = Med(\bar{y}_{itc} \mid c=k+3, t=0)$ for all values of *k*.

Assumption A0 guarantees that the conditional median of [4] yields [5]. Assumption A1 posits that attrition out of FIDE lists occurs only after age 15, and implies that our sample of young players aged 15 is representative of the population of potential professional chess players at the time of entry in the profession. Assumption A2 requires that, conditional on age, median ability does not vary across cohorts. We provide evidence in support of this crucial assumption by looking at ELO scores as proxies of individual ability. Using our auxiliary dataset, we organize individuals in seven 5-yearsage groups, starting from age 25 to 29 and ending with 55 to 59, and consider three points in

[21]For the same reason, median regression is also robust to the presence of a mass of observations with zero productivity documented in Figure 1.

[22]We use E for the mean operator and Med for the median.

time: 2001 (the reference point), 2006, and 2011.[23] For each age group, we compute median ELO scores in 2006 and 2011, and their percentage changes with respect to 2001.Results reported in Figure 3 suggest that these changes have been quite small for each age group, and always below 3 percent. We conclude from this that cohort and time effects are negligible in our data for each selected age group.

The imputation procedure consists of the following six steps:

1) median regressions of productivity y_{itc} on country dummies and the country – specific time and cohort effects X_{it}, to filter out the effects of X_{it} and obtain $\bar{y}_{itc}$;

2) the identification of players belonging to the first cohort c_{15} at time *t=0,*who have dropped out from the sample any time between *t=0* and *t=3*;

3) the imputation to each identified dropout player of a productivity value at time *t=3*arbitrarily larger (smaller) than median productivity at time *t=3* if $\bar{y}_{i0c_{15}} > Med(\bar{y}_{i0c_{15}})$ ($\bar{y}_{i0c_{15}} < Med(\bar{y}_{i0c_{15}})$);[24]

4) the computation of median productivity at the end of the observation period (*t=3*) – or $m_{3,15} = Med(\bar{y}_{i3c_{15}})$ - on the repopulated sample;

5) the computation of $m_{0,18} = Med(\bar{y}_{i0c_{18}})$. If $m_{0,18} > m_{3,15}$, we add to cohort c_{18} as many players (*bots*) with an arbitrary low productivity as required to re-establish $m_{0,18} = m_{3,15}$, which must hold under assumption *A2*. We add *bots* with arbitrarily high productivity if instead $m_{0,18} < m_{3,15}$. By so doing, we repopulate cohort c_{18} at time *t=0* and correct for the self-selection affecting this cohort prior to *t=0*. The repopulated cohort is then used as the basis for the next iteration, involving cohort c_{21}.

6) steps 2-5 are repeated for each cohort until the last, c_{60}.Once the sample has been repopulated, equation [4] can be estimated using median regressions.

We show in the Appendix the results of a Monte Carlo experiment designed to evaluate how our imputation performs in a controlled setting. Results indicate that the procedure is capable of reproducing the medians of the original population rather well. In our method, the assumed lack of

[23]We start from the age group 25-29 because for earlier age groups we only have either one or two observations over time.

[24]We allocate dropouts with productivity close to the median at time *t=0*either above or below the median at time *t=3* as follows. First, we compute the distribution of $D = \bar{y}_{i0c_{15}} - Med(\bar{y}_{i0c_{15}})$ for all the players with $D > 0$ and the distribution of $D' = Med(\bar{y}_{i0c_{15}}) - \bar{y}_{i0c_{15}}$ for the players with $D \leq 0$. Next, we define as close to the median the dropouts with *D* or *D'* smaller than the 10th percentile of the relevant distribution. For these players, we impute an arbitrarily large value of productivity with probability P (resp. 1-P') and an arbitrarily low value of productivity with probability 1-P (resp. P') if $D \leq pct_{10}(D)$ (resp. $D' \leq pct_{10}(D')$),where $pct_{10}(.)$is the 10th percentile of the distribution and $P = \frac{D}{\max(D)}$ and $P' = \frac{D'}{\max(D')}$. To all remaining dropouts we impute an arbitrarily large (small) productivity if $D > pct_{10}(D)$ ($D' > pct_{10}(D')$).

selection for the youngest cohort – our Assumption A1 – trickles down to the older cohorts because of Assumption A2.To illustrate the logic of our imputation strategy, consider the youngest cohort. Under the anchoring assumption, A1, at the end of step 3 the cohort c_{15} at time *t=3* is not self-selected, and median productivity in this re-populated cohort is the counterfactual median that would have been observed had cohort c_{15} not suffered attrition between time *t=0* and *t=3*. Under stationarity (Assumption A2), this median is also the counterfactual median productivity for cohort c_{18} at time *t=0*.Cohort c_{18} at time *t=0* is re-populated with *bots* until its median productivity meets the counterfactual (step 5). These *bots* are considered as dropouts between *t=0* and *t=3* and treated accordingly. Finally, all dropouts (actual dropouts and bots) of cohort c_{18} are used to compute the median productivity that cohort c_{18} would have displayed at time *t=3* had no attrition occurred.

5. Results

We estimate equation [4] on the repopulated sample, using a third order polynomial in age.[25]We report our estimates in the first column of Table 2and plot the associated age-productivity profile in Figure 4. The right panel of the figure reports median productivity by age in the repopulated sample (the dots), after normalizing the median value of productivity at age 15to 1. The left panel of the figure reports instead the normalized age-productivity profile in the unadjusted sample. We estimate that in the repopulated sample median productivity rises from age 15 to peak age 21[26] by slightly more than 5 percent and then declines almost monotonously. Productivity is equal to its initial value at age 15 just after age 30, to 95 percent of the initial value at age 40, and to less than 90 percent of the initial value at age 50.

The comparison with the left panel of the figure points out several important differences: first, relative productivity is always lower in the repopulated sample; second, productivity from baseline to peak age increases much more in the unadjusted than in the repopulated sample (18 percent), suggesting that the increase in the former case is mainly due to the endogenous attrition of less talented players; third, peak age in the repopulated sample occurs earlier than in the unadjusted sample; finally, productivity from peak age declines rather smoothly and at a similar pace in both samples, indicating that the effect of endogenous selection is less important at later ages, in line with our finding that dropout rates decline with age.

As a result of these differences, while in the unadjusted sample productivity at age 50 is about as high as at the baseline age and it is only about 5 percent below by age 60, in the repopulated sample

[25]Adding additional powers to the polynomial does not alter the goodness of fit. Notice also that the variables in the *X* vector are filtered out in the first step of the imputation process.

[26]According to the polynomial estimated in Table 2, peak age is 21.51.

productivity at age 50 is already about 10 percent lower than at the baseline (15 percent by age 60). We conclude that failure to properly address endogenous selection leads to over-estimating productivity at all ages relative to baseline age, and to distorted estimates of the age-productivity profile, that inflate productivity increases until peak age and, to a much lesser extent, productivity declines after the peak.

Columns (2) and (3) of Table 2 report our results when we change the definition of productivity[27]. In column (2) we weight wins and draws with the average ELO score of the opponents met in a given year rather than in each game. In column (3) we consider the raw winning rate, defined as the number of wins plus the number of draws weighted by 0.5 over the total number of games played in a given year. Results turn out to be remarkably similar to our baseline specification in column (1), with marginal changes in the estimated peak age.

Conclusions

We have used data on professional chess tournaments to study the relationship between age and mental productivity in a brain-intensive profession. Using chess has the advantage that individual productivity can be measured with accuracy. We have shown that selective attrition is an important phenomenon, and that the age-productivity profile is heterogeneous with respect to ability. When productivity is not separable in terms of ability and age, the fixed effects estimator produces consistent estimates of the age-productivity profile only when applied to first-differenced data. In this case, however, the linear term of the age polynomial cannot be identified.

We have therefore turned to imputation to repopulate our sample, and to median regressions, using the property that estimates are affected by the position of productivity observations with respect to the median but not by the specific values of imputed productivity. We have compared the estimated age – productivity profiles in the unadjusted and in the repopulated sample, and found that accounting for endogenous selection dramatically reduces productivity at all age, relative to baseline age, and alters significantly the shape of the age-productivity profile. While in the unadjusted sample median productivity at age 50 is about as high as at age 15,in the repopulated sample it is about 10 percent lower. We have argued that productivity is higher in the unadjusted sample because weaker players are more likely to dropout from chess as they age. The effect of selection is particularly pronounced at earlier ages, and generates a steep increase in productivity until peak age. Because of this large increase, productivity at later ages remains above baseline productivity, in spite of an almost linear decline after the peak.

Our results highlight the importance of accounting for endogenous selection, and do not confirm the recent results by Van Ours, 2009,suggesting that for economists, another brain-intensive

[27]The size of the repopulated sample changes across columns in the table because the number of *bots* required to perform the imputation varies with the measure of productivity being used.

profession, mental productivity does not decline with age. We find that chess professionals aged 40 are about 5 percent less productive than young players aged 15.

This paper is not alone in emphasizing that productivity is not separable in terms of age and ability. In the Mincerian tradition, it has been customary to assume that earnings (and productivity) are separable in age (experience) and education (ability), mainly because of lack of data on lifetime earnings. Recent work by Heckman, Lochner and Todd, 2008, however, has shown that this assumption is convenient but not supported by the empirical evidence.

References

Bloom David E. and Alfonso Souza-Poza, 2013, Ageing and Productivity: Introduction, *Labour Economics*, Special Issue.

Borsch-Supan Axel and Matthias Weiss, 2007, Productivity and Age: Evidence from Work Teams at the Assembly Line, MEA Working Paper 148

Castellucci Fabrizio, Padula Mario and Giovanni Pica, 2011, The Age-Productivity Gradient: Evidence from a Sample of F1 Drivers, *Labour Economics*, 18, 464-473

Glickman Mark, 1995, A Comprehensive Guide to Chess Ratings, *American Chess Journal*, 3, 59-102

Göbel Christian and Thomas Zwick, 2012. Age and Productivity: Sector Differences, *De Economist*, 160(1), 35-57.

Gransmark Patrik and Christer Gërdes, 2010, Strategic Behaviour across Gender: A Comparison of Female and Male Expert Chess Players*, Labour Economics*, 17, 766-775

Heckman James J., Lance J. Lochner and Petra E. Todd, 2008, Earnings Functions and Rates ofReturn, *Journal of Human Capital*, 2(1), 1-31

Imbens Guido and Jeffrey Wooldridge, 2007, *What's New in Econometrics: Linear Panel Data Models.* NBER Summer Institute 2007.

Myck Michal, 2007. Wages and Ageing: Is There Evidence for the 'Inverse-U' Profile? *Oxford Bulletin of Economics and Statistics*, 8,202-229.

Olivetti Claudia and Barbara Petrongolo, 2008, Unequal Pay or Unequal Employment? A Cross-country Analysis of Gender Gaps, *Journal of Labor Economics*, 621-654

Pischke, Jörn-Steffen, 2001, Continuous Training in Germany, *Journal of Population Economics*, 14, 523-48

Pekkarinen Tuomas, and Roope Uusitalo, 2012. Aging and Productivity: Evidence from Piece Rates. IZA Discussion Paper 6909

Skirbekk Vegard, 2003, Age and Individual Productivity: A Literature Survey, Max Planck Institute for Demographic Research Working Paper

Thurstone Louis Leon, 1927, A Law of Comparative Judgement, *Psychological Review*, 34, 273-286

The Economist, 2004, Over 30 and Over the Hill, June 24th

Van Ours Jan, 2009, Will You Still Need Me – When I'm 64?, IZA Discussion Paper 4246

Weinberg Bruce A. and David W. Galenson, 2005. Creative Careers: The Life Cycles of Nobel Laureates in Economics, NBER Working Papers 11799.

Tables and Figures

Table 1. Descriptive Statistics

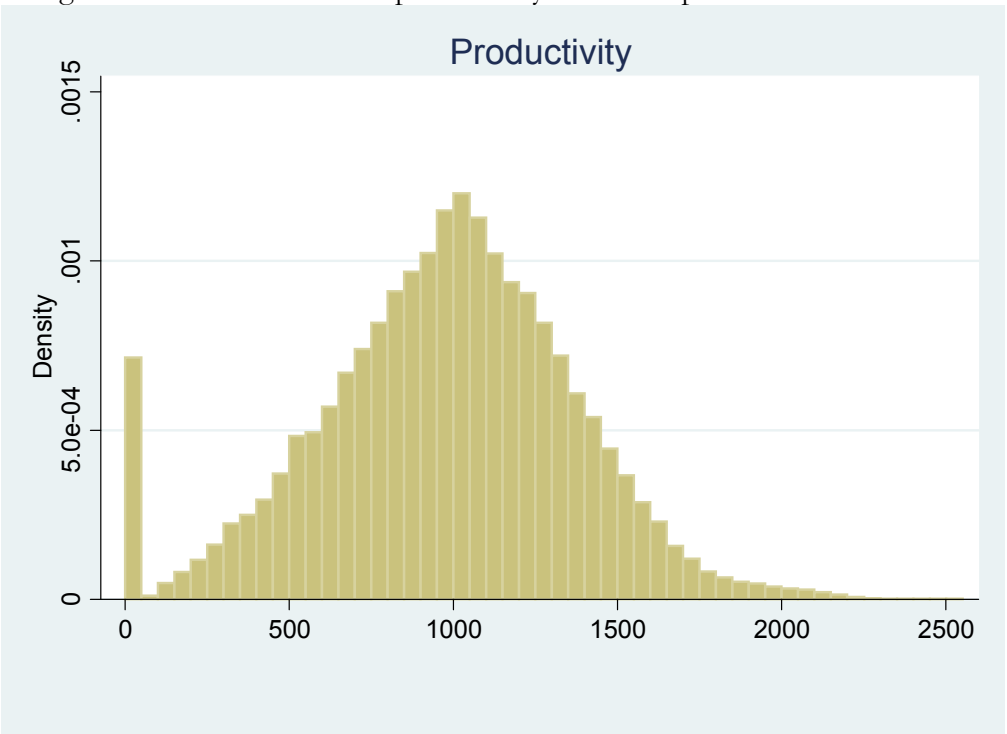
Variable	Mean	Std. Dev.	Min	Max
Age	38.09	12.31	15	60
Games	17.45	19.20	1	289
Productivity	972.39	396.66	0	2551
GDP per capita (in thousand $ at constant prices)	22.06	12.01	2.64	53.48
Internet users (per 100 inhabitants)	57.19	23.05	4.38	96.62
Number of tournaments	139.8	156.97	4	674
GDP per capita growth at age 15-25	2.15	2.29	7.15	21.97
Average number of internet users at age 15-25 (per 100 inhabitants)	14.09	23.11	0	92.60

Table 2 – Estimates of Eq. [4]using median regressions on the repopulated sample. Dependent variable: alternative measures of productivity

	Baseline	**With average ELO of opponents as weight**	**Un-weighted winning rate**
Age	38.98***	37.21**	0.02***
	(7.01)	(6.88)	(0.002)
Age^2/100	-126.00***	-121.70***	-0.06***
	(20.35)	(19.98)	(0.008)
Age^3/1000	10.97***	10.74***	0.005***
	(1.85)	(1.82)	(0.001)
Peak age	21.51	21.28	22.27
Number of players in the unadjusted sample	40,545	40,545	40,545
Number of players in the repopulated sample	52,385	52,226	50,041

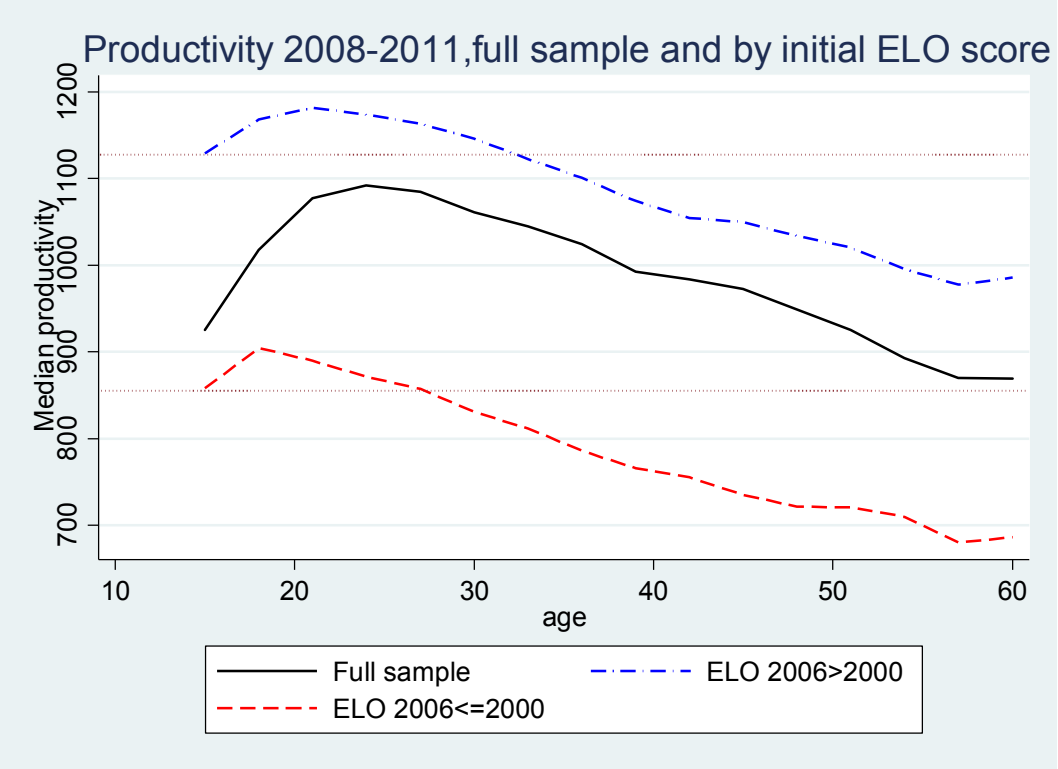
Note: country dummies, real GDP per capita, the number of tournaments, the percentage of internet users per 100 inhabitants, real GDP growth and growth of access to internet when individuals were aged between 15 and 25 have been filtered from the data in the first step of the imputation process. Three, two and one star for statistically significant coefficients at the 1, 5 and 10% level of confidence.

Figure 1. The distribution of productivity in the sample



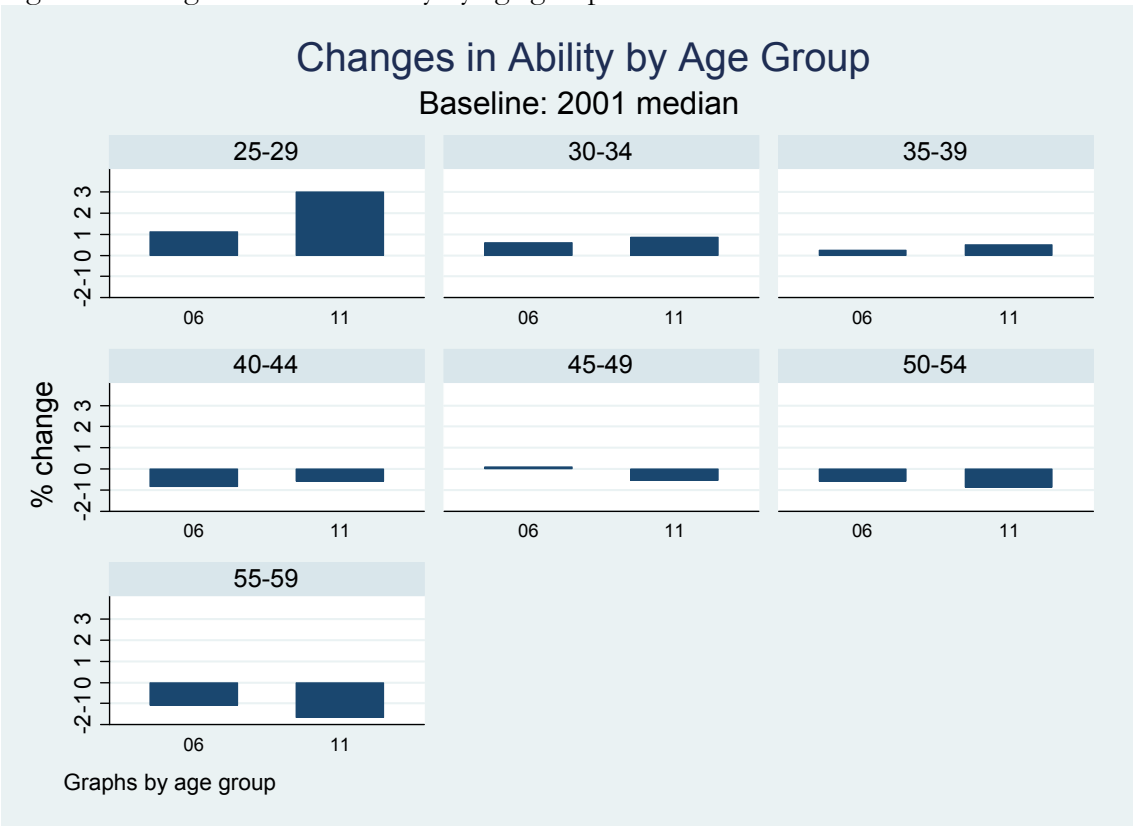
Source: main dataset

Figure 2. Median productivity by age, full sample and by initial ELO score.



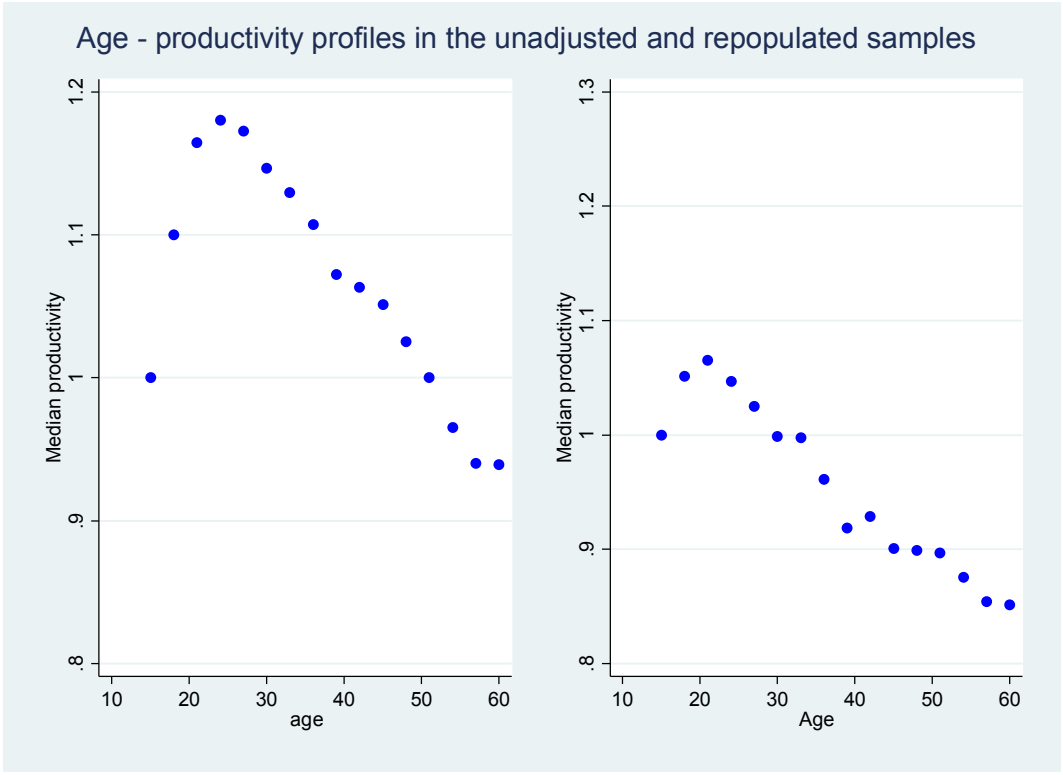
Source: main dataset

Figure 3. Changes in median ability by age group



Source: auxiliary dataset

Figure 4. Median age-productivity profiles in the unadjusted and in the repopulated sample.



Source: main dataset

Appendix

We perform a Monte Carlo experiment to evaluate how well our imputation strategy reproduces the initial population starting from a selected sample. Suppose that the data generating process is given by

$$y_{itc} = 5 + 40A_{itc} - 120\frac{A_{itc}^2}{10^2} + 10\frac{A_{itc}^3}{10^3} + 20\,\alpha_i + \alpha_i A_{itc} + \\ +10\,CCO_{ic} + 2\,CP_{ct} + 10\,P_2 - 50\,P_5 + 60\,P_{10} + \varepsilon_{itc} \qquad [A1]$$

where individual talent is $\alpha_i \sim N(0,1)$ and the noise term ε_{itc} is distributed as *N(0,10)*. The country-specific cohort and period effects are CCO and CP respectively, while P are country dummies. There are 16 cohorts, {15,18,21,...,60}, each composed of 3,000 individuals observed at time *t=0* and at time *t=3*, and 10 countries. Players are ranked on the basis of their name and systematically allocated to countries (the first player is allocated to the first country, the second player to the second country, the eleventh player to the first country again and so on). We assume that that country fixed effects are all zero except for countries 2, 5 and 10.

The initial sample is composed of 96,000 players. Given the features of the data generating process, ability is orthogonal to age in the population. Players self-select into chess on a year-by-year basis according to a selection rule that depends on age, ability, and an individual-by-time specific random shock v_{it}, that follows a N(0,0.1) distribution. The selection rule is defined as follows:

$$Z_{itc} = 1 \quad \text{if} A_{itc} = 15$$

$$Z_{itc} = 1 \quad \text{if } \alpha_i + v_{itc} > 0.035A_{itc} - 2.5 \text{ and } A_{itc} > 15 \qquad [A2]$$

$$Z_{itc} = 0 \quad \text{otherwise}$$

where Z is an indicator that takes value 1 when player *i* keeps playing and 0 when he drops out. After selection, the correlation between age and talent is assumed to be equal to 0.187, and average ability increases with age. The main feature of this selection process is that less talented players drop out as they get older, coherently with our estimates.

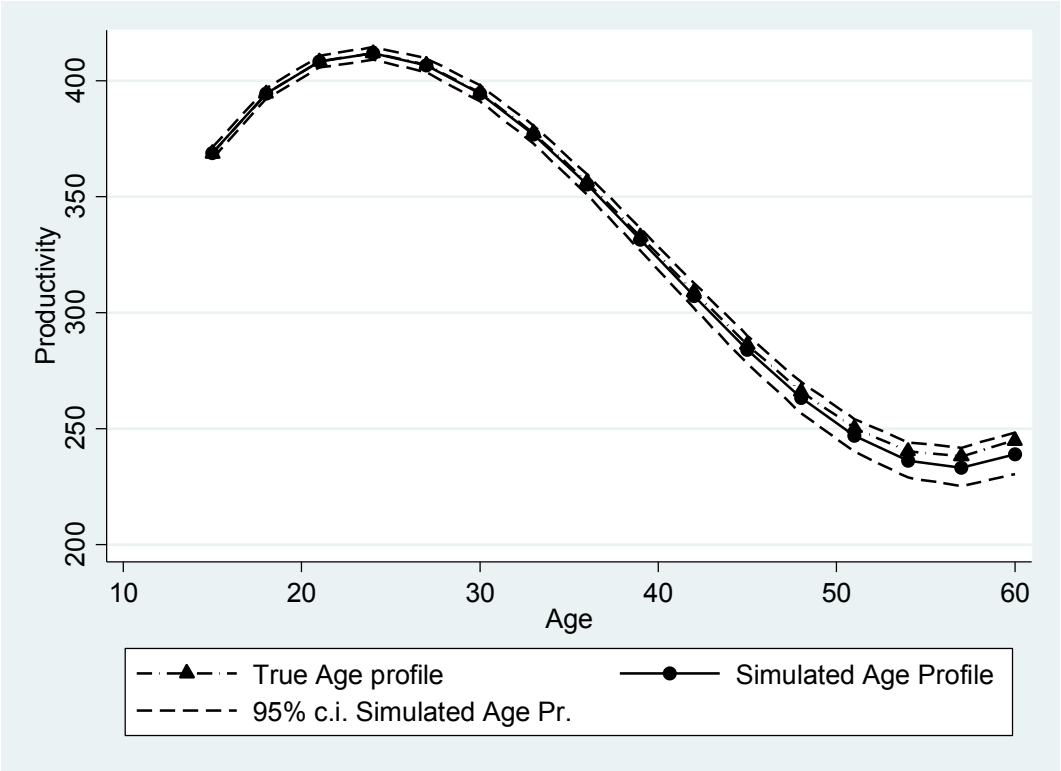
The simulation operates as follows: after randomly drawing the initial population, we apply our selection rule to generate attrition and obtain a selected sample. We then apply to this sample the imputation procedure described in the text. Next, we estimate the following model on the repopulated sample

$$Med(\bar{y}_{itc}|A_{itc}) = \beta_0 + \beta_1 A_{itc} + \beta_2 A_{itc}^2 + \beta_3 A_{itc}^3 + \xi_{itc} \qquad [A3]$$

where, as in the main text, $\bar{y}_{itc}$ represents productivity net of country specific cohort and period effects. We repeat this procedure 500 times. The estimates obtained from (A3) are shown in Figure A1, where we plot the age-productivity profile obtained from our Monte Carlo simulation, its 95%

confidence interval, and the profile from the data generating process. The two overlap almost perfectly and the confidence interval is remarkably narrow.

Figure A1. True and Simulated Age-Productivity Profiles – Monte Carlo Analysis



Chapter 3

Laterborns Don't Give Up. The Effects of Birth Order on Earnings in Europe*

by

Marco Bertoni
(University of Padova and CEP, LSE)

Giorgio Brunello
(University of Padova, IZA and CEsifo)

Abstract

While it is well known that birth order affects educational attainment, less is known about its effects on earnings. Using data from eleven European countries for males born between 1935 and 1956, we show that firstborns enjoy on average a 13.7 percent premium over laterborns in their wage at labour market entry. However, this advantage is short lived, and disappears by age 30, between 10 and 15 years after labour market entry. While firstborns start with a better match, partly because of their higher education, laterborns quickly catch up by switching earlier and more frequently to better paying jobs. We argue that a key factor driving our findings is that laterborns are more likely to engage in risky behaviours.

Key words: birth order, earnings, risk aversion, Europe
JEL Code: D13, J12, J24

* We thank Hessel Oosterbeek, Erik Plug, Olmo Silva, Guglielmo Weber, Christoph Weiss and the audience at workshops in Reus, Padova and Rome (Brucchi Luchino) for comments and suggestions. This paper uses data from SHARELIFE release 1, as of November 24th 2010 and SHARE release 2.5.0, as of May 24th 2011. The SHARE data collection has been primarily funded by the European Commission through the 5th frame work 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 frame work programme (SHARE-PREP, 211909 and SHARE-LEAP, 227822).

Introduction

Does birth order affect wages? According to Ruth Mantell of the Wall Street Journal, 2011, the answer is clearly positive. She reports that firstborn kids are "...the most likely to earn six figures and hold up a top executive position among workers with siblings...". She also quotes economist Sandra Black as saying that "... birth order affects educational attainment, which then affects earnings […]. Laterborns earn less than firstborns, and a substantial part of this difference is due to the fact that laterborns get fewer years of education."

While there is substantial empirical research investigating the effects of birth order on educational attainment, less has been done to explore the effects on earnings. One reason could be the scarcity of datasets containing information both on earnings and on birth order. Another reason, we suspect, is that the research question is viewed as not particularly interesting. If one believes, as many economists do, that earnings are a function of human capital, evidence that firstborns have better education implies that they also have higher earnings.

Yet – with a single important exception[1] - the few studies that have addressed this issue have found that the effects of birth order on earnings are rather negligible, in spite of the significant effects on educational attainment. For example, in their study of Swedish data, Björklund and Jäntti, 2012, find that firstborns attain on average 0.2 more years of education than laterborns, but only a 0.25% premium on earnings between ages 31-40. Given that returns to education in Sweden range between 3.5 and 5.5 percent (see Harmon, Oosterbeek and Walker, 2003), the estimated premium is much lower than the expected 0.7-1.1 percent. These studies also focus on earnings at a given point in a working lifetime, typically before age forty, or on average earnings over a short period of the working life cycle, and are therefore silent on whether birth order has a temporary or a permanent impact on individual earnings.

In this paper, we contribute to this small literature by studying the effects of birth order on earnings over the life cycle in a sample of 4,280 males born between 1935 and 1956 and residing in eleven European countries (Austria, Belgium, the Czech Republic, Denmark, France, Germany, Italy, the Netherlands, Spain, Sweden and Switzerland). We consider several measures of real annual earnings: the entry wage - defined as the initial wage in the first job - wages at age 30, 40 and 50, and the current or last wage, defined either as the wage in the job currently held if still active at age 50 plus or as the wage in the last job before retiring. We also add a measure of lifetime earnings, or the discounted value of the stream of earnings from age ten to retirement. By looking at earnings at different points in the life cycle, and at lifetime earnings, we can tell whether the estimated birth order effects on earnings are temporary or permanent.

[1] Kantarevic and Mechoulan, 2005, find that order of birth has a significant effect on hourly earnings in a relatively small sample of US workers.

We show that the advantage enjoyed by firstborns over laterborns is short lived: they earn on average a 13.7% premium in their entry wage, but this advantage is completely gone by age 30. We also find that being a firstborn has no statistically significant effect on earnings at age 50 and on the current wage. Since the initial wage gains are quickly lost, and laterborns start working earlier than firstborns, it is not surprising that being a firstborn has no statistically significant effect on lifetime earnings.

The temporary advantage enjoyed by firstborns implies that birth order has a positive effect on earnings growth, measured as wages at age *t* minus the entry wage. Importantly, we find that this effect remains even after controlling for educational attainment. This suggests that differences in education between firstborns and laterborns are not sufficient to explain the observed differences in wages over the lifecycle. We also find that education negatively affects earnings growth, a result consistent both with the learning model by Altonji and Pierret, 2001, and with the human capital model, provided that education and experience are substitutes in the production of skills.

Temporary birth order effects are closely associated to differences in job-to-job mobility after labour market entry. On the one hand, firstborns find better initial matches – not only they earn more, but they are also more likely than laterborns to be in white collar and in public sector jobs - and stay on their initial jobs longer. On the other hand, laterborns start with poorer matches but change jobs swiftly, and by virtue of job mobility quickly catch up with firstborns. To illustrate the effects of mobility, we compare expected log wages at age 30 for firstborns and laterborns and find that they are quite similar (a 0.7% advantage for laterborns). These wages can be expressed as the weighted average of log wages for those still in the first job at age 30 and log wages for those in other jobs, using as weights the probability of being in the first job at age 30. While firstborns who are still in their first job at age 30 retain a 5% advantage on earnings over laterborns in their first job at 30, this advantage is more than compensated by the fact that, at that age, laterborns have a higher probability of being already in their second or third job, that pay higher earnings than the first job. A similar pattern holds at age 40 as well.

Drawing on a vast literature in psychology (see for instance Sulloway, 2007) and using our own evidence in support, we argue that firstborns differ from laterborns both because they have higher education and because they are less likely to engage in risky behaviours (see Wang et al., 2009). On the one hand, better education explains why firstborns start with a better match. On the other hand, the higher propensity to take risks explains why laterborns incur in higher turnover (see Allen et al, 2005) and enjoy higher wage growth than firstborns (see Shaw, 1996).

The paper is organized as follows: we briefly review the relevant literature in Section 1, introduce the data in Section 2 and discuss the empirical methodology in Section 3. Our results are reported in Section 4. We discuss our findings in Section 5 and present a few extensions in Section 6. Conclusions follow.

1. Review of the Literature

The effects of birth order on educational attainment have been widely studied. In a recent influential contribution, Black, Devereux and Salvanes, 2005, (BDS from now on) use Norwegian registry data and find that birth order has a significant and large negative effect on children's education, even after controlling for family size. In particular, they estimate that being a second child reduces educational attainment with respect to being a firstborn by close to 0.3 years of schooling. Negative effects have been found also in recent research by Bagger et al., 2013, for Denmark, Björklund and Jantti, 2012, for Sweden, and De Haan, 2010, and Kantarevic and Mechoulan, 2005, for the US.

Less has been done to investigate the effects of birth order on earnings. Most of the existing studies are based on US data and consider earnings relatively early in an individual's career (before age 40). While results are sensitive to the inclusion of covariates, the broad assessment is that the estimated effects tend to be small or negligible. Behrman and Taubman, 1986, use US data for young adults and show that, after adjusting for age or work experience, there are differences by birth order in both schooling and log earnings. The effects on earnings, however, become statistically insignificant when they include controls for observed childhood family background characteristics.

Olneck and Bills, 1979, examine the effect of birth order and family size on childhood test scores and adult levels of education, occupation, and wages, finding a negligible influence of birth order on all measures of achievement. Kessler, 1991, uses data from the US National Longitudinal Survey of Youth to examine the effect of birth order and family size on individual behaviour over the course of teenage and early adult lives. He finds that neither birth order nor childhood family size significantly influences the level or growth rate of wages for individuals aged 14-22, 18-26 and 22-30. Björklund and Jäntti, 2012, use Swedish registry data and report that the firstborn child attains 0.2 years of additional education and earns around 0.25% higher long-run earnings (earnings are measured at ages 31-40) than other siblings. After examining other outcomes, they conclude that birth order is not a major source of the family impact on economic outcomes and thus not a major source of inequality of opportunity.[2]

To our knowledge, the work by Kantarevic and Mechoulan, 2005, stands out as the only paper to date that finds significant effects of birth order on (hourly) earnings. The authors use data from the Childbirth and Adoption History File (CAHF), a special supplemental file of the US PSID (Panel Study on Income Dynamics)[3] and find that, when the age of the mother at birth is omitted from the vector of covariates, birth order has no statistically significant effect on earnings. When age is

[2] Yet, in a recent contribution, De Haan, Plug and Romero, 2012, find that birth order affects early outcomes in Ecuador.

[3] Their sample is rather small (3000 observations) and pools together males and females.

included, they report that the hourly earnings of firstborns are 6.3% higher than those of laterborns.[4]

2. The Data

In this paper, we use the Survey of Health, Ageing and Retirement in Europe (SHARE), a multidisciplinary and cross-national European data set containing current and retrospective information on labour market activity, retirement, health and socioeconomic status of more than 25,000 individuals aged 50 or older. We draw our data from the first three waves of the survey, and in particular the third wave, SHARELIFE, which contains detailed retrospective data on life and labour market histories. We focus on males because of the problems associated with female labour force participation and exclude the self-employed and people aged 50+ who have worked less than 5 years.[5] In SHARELIFE, survey participants are asked to report the amount they were paid monthly after taxes each time they started an employment spell. They are also asked the monthly net wage in their current job (if they are still working) and the monthly net wage at the end of the main job in their career (if they have already retired). For wages and other benefits to be comparable across time and country, we follow Brunello, Weber and Weiss, 2012, and transform them into 2006 Euro using PPP exchange rates and CPI indices.

We use these rich data to construct for each individual several measures of real annual earnings, that span his working life from the first to the current or last job. We start with the entry wage W_1, defined as the initial wage in the first job. Since information on this wage is missing for about 25 percent of the individuals in our final sample, we use predictive mean matching to impute missing data and obtain $\overline{W_1}$.[6] We also compute the initial wage in the second and third job, the current or last wage, the wages at age 30, 40 and 50 and lifetime earnings.

We define lifetime earnings (or permanent income) as the income flowing from the asset value of working at age ten. The construction of this variable and of wages at different ages is described in detail both in Appendix A of Brunello, Weber and Weiss, 2012, and in Weiss, 2012. In short, for those who have had only one job in their working life (more than 20 percent of the sample), we interpolate between the first wage and the last (or current) wage. For those who have had more than one job, we observe the first wage in each job as well as the current or last wage. For this second group, we regress current wages on labour market experience, a rich set of controls, which include

[4] The statistical significance of this effect falls from 5 to 10% when father's education and the age of the father at childbirth are added to the covariates.

[5] Murphy and Welch, 1990, also exclude the self-employed in their analysis of age-earnings profiles.

[6] As shown in Section 4, our results do not depend on imputation. Predictive mean matching replaces a missing value with the observed value for which the predicted value is the closest to that of the missing value. See Weiss, 2012, for details. The percentage of missing values is very similar among firstborns (23.2%) and laterborns (24.3%).

education, occupation, sector of activity, cohort and country effects and economic conditions at age ten, and the interactions of these controls with experience. We then use the estimated coefficients and the first wage in each job to generate both the final wage in the job and within-job earnings growth.[7] With this information in hand, we compute annual wages at age 30, 40 and 50 and the discounted value of earnings at age ten, using a 2 percent discount rate. [8]

Our dataset has the advantage that it covers eleven European countries, and the potential drawback that it uses long recall data. These data are subject to measurement error, possibly not of the classical type. However, as discussed in Brunello, Weber and Weiss, 2012, validation studies have found that recall bias is not severe in SHARELIFE data, arguably because of the state-of-the-art elicitation methods used: respondents are helped to locate events along the time line, starting from domains that are more easily remembered, and then asked progressively more details about them.[9]

Our final sample consists of 4,280 males born between 1935 and 1956 and residing in Austria, Belgium, the Czech Republic, Denmark, France, Germany, Italy, the Netherlands, Spain, Sweden and Switzerland.[10] While waves 1 and 2 of the survey have information on order of birth ("Were you the oldest child, the youngest child, or somewhere in-between?"), wave 3 has data on individual and household conditions at age ten. We rely on answers to the question "Including yourself, how many people lived in your household at this accommodation when you were ten?" to measure gross family size, which includes both siblings and other members.[11] We also use the answers to the question "Who lived in the household when ten" to estimate net family size, or the number of siblings, by subtracting other members (parents, grandparents and other relatives) from gross family size. In our data, the average household size at age ten is 5.44 members, and the average number of siblings is 3.34.[12] As shown in Table 1, the distribution of siblings varies with whether the interviewed individual is the oldest child or not, mainly because 24 percent of oldest children are only children. Compared with the distribution of siblings in the Norwegian sample used by BDS, our sample comprises households with a higher number of siblings, which reflects both the different sample period – the individuals are born between 1935 and 1956 in our sample and between 1912 and 1984 in BDS's sample - and the fact that our sample includes also Southern European countries, where

[7] Brunello, Weber and Weiss, 2012, show that estimates are broadly unaffected when they replace labour market experience with age and exclude education in the wage regressions used to generate both the end wage in each job and within-job earnings growth for individuals who have had more than one job.

[8] We are very grateful to Christoph Weiss for providing the codes required to compute earnings profiles and lifetime earnings from the third wave of the survey SHARE.

[9] Brunello, Weber and Weiss, 2012, validate this procedure by comparing predicted and actual wages in the German GSOEP and find that predictions based on the methodology suggested in the text are accurate.

[10] By selecting only individuals born from 1935 onwards, we reduce the role of survivorship bias (see Modin, 2002) and recall bias for older workers, the weight of imputation, and also make sure that no individual in our sample entered the labour market before the second World War.

[11] Needless to say, household size at age ten is less correlated with order of birth than household size at birth. For the small minority of individuals for which this information was not available – around 2 percent of our sample – we reconstruct sibship size using information on the number of siblings alive at the time of the first SHARE interview.

[12] We recode the number of siblings so that the top category is 10 or more.

the number of siblings is typically higher (2.90 in Sweden and 3.88 in Spain).

The third wave of SHARE also contains a wealth of data on household and individual conditions at age ten. We define the vector *X* as comprising the following covariates: whether the household was located in a rural area or a village, dummies for the profession of the main breadwinner, a dummy for the presence of hunger episodes before age 15, a dummy indicating whether parents smoked, drank heavily o had mental health problems during childhood, a dummy if one parent died before age 35, and dummies for the presence of parents, grandparents or foster parents in the household.[13]

Unfortunately, our information on the age of the parent at birth is available only for those parents who were still alive at the time of the interview. We check whether omitting this critical piece of information significantly affects our estimates by running our regressions with and without the age of the mother at birth in the sub-sample where this measure is available. As reported below, our evidence suggests that omitting maternal age at birth does not affect our estimates in a qualitative way.

Table 2 shows the summary statistics of the main variables used in this study, separately by order of birth (firstborns and laterborns). The statistics for the full sample are reported in Table A1 in the Appendix. These tables suggest that firstborns are on average better educated than laterborns (12.59 versus 11.49 years of schooling), start working later (at age 19.6 versus 18.6) and have a substantially higher entry wage (11,786 real euro versus 10,577, a 11.4% premium). This "premium" declines with the second and third job and with age and is close to 3.3% in the current or last wage (23,546 versus 22,787). Firstborns have fewer siblings (1.51 versus 2.91) than laterborns. Furthermore, the households where firstborns lived at age ten were more likely to be located in urban areas and to have a white collar breadwinner, indicating that household wealth was also higher.

3. Empirical Methodology

We estimate the following linear regression model:

$$\ln w_{it} = \alpha + \beta O_i + \gamma F_j + \delta X_i + \mu_s + \mu_c + \varepsilon_{it} \qquad (1)$$

where the subscripts *i, j* and *t* are for individuals, households and time, *w* is annual real earnings, *O* is a dummy equal to 1 if the individual is firstborn and to 0 otherwise[14], *F* is the number of siblings in the household when the individual was ten, the vector *X* is described in the previous section, μ_c and μ_s are cohort and country fixed effects and ε_{it} is an error term, which can be decomposed as

[13] We exclude information such as the number of books in the household and housing facilities at age ten because they could be affected by birth order, as suggested by De Haan, Plug and Romero, 2012.

[14] As in BDS, we treat children without siblings as firstborns. As discussed later in the paper, sensitivity analysis which excludes firstborns yields very similar results.

$\varepsilon_{it} = \lambda_j + \eta_i + v_{it}$, where λ_j and η_i are family and individual fixed effects and v is random noise. Since we are interested in the effects of being firstborn on earnings at different points of the life cycle, we use as dependent variable (in logs): the entry wage, the initial wage in the second and third job, the wage at ages 30, 40 and 50, the current or last wage and lifetime earnings.

As discussed by Bagger et al, 2013, family size can be viewed as the outcome of inter-temporal utility maximization by altruistic parents, and the family fixed effect λ_j as a function of parental spending and preferences, partly unobserved by the analyst. Parental choice implies that family size is a function of λ_j. Since parents typically choose size and individual investment in human capital, which affects earnings, the family fixed effect influences individual outcomes directly. Birth order, on the other hand, depends directly on family size and only indirectly on the family fixed effect.

The identification of birth order effects in Eq. (1) is complicated by the fact that, while the order of birth may well be considered as good as randomly assigned within a given family, the question is less clear-cut when variation between families is also used, as we do. As shown in Table 2, firstborn individuals belong more frequently to smaller families, and smaller families are not only typically better off, but may also devote more time and economic resources to each child (the quality-quantity trade-off discussed by Becker and Lewis, 1973). Since family size depends both on observable and on unobservable parental traits that may also be related to earnings capacity, the omission of some of these traits in Eq.(1) biases the estimated coefficient of family size, and contaminates the estimates of birth order effects.

BDS address this problem by using two approaches: the first approach relies on selection on observables and consists of including a rich set of covariates describing economic and social conditions of families, in the hope that this set mops up the family fixed effect. In the second approach, they use family fixed effects, thereby focusing on within-family variation in educational outcomes. We capture some household traits by conditioning our estimates on the covariates included in vector *X*. When these effects are netted out and we estimate (1) by ordinary least squares, the bias in the estimated coefficient of birth order is

$$\beta_{OLS} = \beta + \frac{Cov(O_i, F_j)}{Var(O_i)}(\gamma - \gamma_{OLS}) + \frac{Cov(O_i, \lambda_j)}{Var(O_i)} + \frac{Cov(O_i, \eta_i)}{Var(O_i)} \qquad (2)$$

Since birth order depends on λ_j only indirectly, $Cov(O_i, \lambda_j) = 0$. Furthermore, $Cov(O_i, \eta_i) = 0$ if there are no genes for being firstborn.[15] Therefore, the bias in (2) is driven by the negative

15 BDS, 2005, argue that "…in general, there are no genes for being a firstborn or a laterborn so it is unlikely that the birth order effects we find have genetic or biological causes...", p.20. De Haan, Plug and Rosero, 2012, have recently questioned this assumption on the ground that that laterborns may face higher prenatal environmental risks because of increased levels of maternal antibody, that may attack the development of the brain in utero.

correlation between order of birth and family size $Cov(O_i, F_j)$ and by the OLS bias in estimated family size effects $(\gamma_{OLS} - \gamma)$. By removing this bias, family fixed effects guarantee that the estimate of birth order effects is consistent. Alternatively, one can set to zero the covariance between order of birth and family size by estimating separate regressions by family size, as done for instance by BDS.

Since in our data we do not observe multiple members within the same original family, we cannot estimate (1) using family fixed effects. We therefore estimate Eq. (1) by family size and show that the qualitative results based on these estimates are broadly unaffected when we pool different family sizes. This suggests that the bias induced by pooling has relatively small effects on the coefficient of interest, which measures the effects of birth order on labour market outcomes. Reassuringly for our estimation strategy, BDS find that birth order effects on educational attainment are rather homogeneous across families of different size, and that their estimates do not vary much when family fixed effects are added to tease out unobservable family characteristics.

Notice that empirical strategies that rely on family fixed effects are not entirely free of problems. To see why, consider that within a given family firstborn and laterborn children usually belong to different birth cohorts, and therefore tend to face different macroeconomic and labour market conditions at several key moments of their lives.[16] This may confound the effect of birth order on earnings.

Since we have measures of real annual earnings at different points of an individual working life as well as a measure of lifetime earnings, we can study how the effects of birth order on earnings vary over the life cycle. To illustrate, suppose that firstborns have a higher initial wage in their first job than laterborns, and assume that we can observe the wage of both groups at age 50. We can then estimate

$$lnW_{i50} - lnW_{iF} = (\alpha_{50} - \alpha_F) + (\beta_{50} - \beta_F)O_i + (\gamma_{50} - \gamma_F)F_j + (\delta_{50} - \delta_F)X_i + \phi_s + \phi_c + (v_{i50} - v_{iF}) \quad (3)$$

where the subscripts *50* and *F* are for the late and the entry wage, and the parameters ϕ are country and cohort effects. This approach has the advantage that it differences out both family and individual fixed effects. Assuming that $\beta_F > 0$, by estimating (3) we can evaluate whether the positive effect of birth order on earnings persists $(\beta_{50} - \beta_F = 0)$, increases $(\beta_{50} - \beta_F > 0)$ or declines $(\beta_{50} - \beta_F < 0)$ over time.

[16] For instance, Angelini and Mierau, 2012, find negative effects of bad macroeconomic conditions at birth on childhood health. Giuliano and Spilimbergo, 2009, estimate negative effects of recessions during early adulthood on self-confidence, locus of control and other beliefs. Lindeboom et al., 2006, find negative mortality effects of a recession at birth. Most relevant for our purposes, Oreopoulos et al., 2010, find negative effects of graduating during a recession on employment and earnings – especially in the short run.

4. Main Results

We introduce the presentation of our estimates by showing in Table 3 the estimated effect of the dummy "oldest child" on educational attainment, both by family size (two, three and four siblings) and by pooling all sizes. We find a positive and statistically significant effect, that ranges between 0.645 and 0.749 years of education, similar to the average effect estimated by BDS for Norway (0.656)[17] but much higher than the effect estimated by Björklund and Jäntti for Sweden (0.248).

Our key results are presented in Table 4, where we show estimated birth order effects both on the entry and on the current or last wage, separately by number of siblings[18] (2, 3 or 4 siblings) and by pooling together all different family sizes, after controlling for sibship size. The table is organized in eight columns, four for each definition of earnings. We find that the dummy "oldest child" has a positive, sizeable and statistically significant effect on the entry wage. Depending on the number of siblings, our estimates suggest that firstborns earn at labour market entry approximately 13.5 to 18.6% more than laterborns, a substantial amount. Yet, this gain is gone by age 50 or later.

The table also shows that our qualitative results are not affected if we pool families with different number of siblings. For instance, we estimate that firstborns enjoy a 13.7% premium with respect to laterborns in their entry wage and no premium at all in their current wage. Because of this, we will focus the presentation of our results in the rest of this section on the sample that pools all family sizes.[19] In Table 5, we look at earnings measured at different points of the lifecycle (age 30, 40 and 50), as well as at lifetime earnings, and confirm that order of birth matters only at labour market entry.

Since some of the data have been imputed, one may worry that our findings are driven by imputation. Table 6 compares estimated birth order effects on entry and current earnings in the samples with and without imputation, and shows that these effects are quite similar. Without imputed data, the marginal effect on the entry wage is slightly smaller than with imputation (12.9% vs. 13.7%). However, the two estimates are not statistically different.

An additional source of concern is that the estimates in Table 4 do not control for the age of the mother at birth. This can affect our estimates, as parents of firstborns are likely to be younger than parents of laterborns. Unfortunately, our data include information on the age of parents at birth only for the interviewed individuals whose parents were still alive at the time of the survey. Given that the survey focuses on individuals aged 50+, this is only a minority of the original sample. Nonetheless, for this smaller sample we can compare estimates with and without controlling for the

[17] This effect is computed as the arithmetic mean of the effect of being the second to the tenth child. See Table 8, column 1 of BDS.

[18] Similarly to Price, 2008, we stop at 4 siblings because sample size would fall drastically if we were to consider households with a higher number of siblings.

[19] Detailed results by family size are available from the authors upon request.

age of the mother at birth. As reported in Table 7, including the age of the mother at birth as additional covariate in the regressions has virtually no effect on our estimates.

Our results suggest that the effect of being firstborn on earnings is temporary and dies out as individuals increase their experience in the labour market. To confirm this, Table 8 presents the estimated effects of birth order on earnings growth over the life cycle, measured alternatively as the difference between earnings at 30, 40, 50 or current earnings and the entry wage. By differencing individual wages over the life cycle, we are able to purge our estimates from fixed family and individual effects. In all cases, the estimated coefficient associated to being firstborn is negative, statistically significant and between -13.5 and -16.2%, confirming that firstborns may have an early advantage, but that laterborns quickly catch up.

We investigate whether the birth order effect disappears when we control for differences in educational attainment by adding years of schooling as an additional covariate in the earnings growth regressions, where the fixed individual and family effects which correlate with education have been removed. Table 9 shows that education attracts a negative and statistically significant coefficient, and that the effect of birth order remains even after conditioning on education, although with a lower absolute value. This finding suggests that education is not the only "mediator" of the effects of birth order on earnings.

Finally, we consider the effects of birth order on the probability of being without a job at different ages, starting with age 20. We estimate linear probability models and report in Table A2 that firstborns have a higher probability of being without a job early on in their career (at age 20 and 25) and are as likely as laterborns to be employed at later ages. Clearly, this effect reflects the fact that firstborns are more likely to stay in school longer.

5. Discussion

Our key findings can be summarized as follows: a) birth order effects on earnings are temporary and decline with labour market experience; b) the effects of birth order on earnings cannot be fully explained by the higher educational attainment of firstborns; c) higher education and earnings growth are negatively correlated. The last result is consistent both with the learning model of Altonji and Pierret, 2001, and with the human capital model if early and later learning episodes are substitutes rather than complements.[20] These models, however, need to be adequately adapted to encompass also findings a) and b).

[20] In their classical paper on employer learning and wage dynamics, Altonji and Pierret, 2001, have shown that, when employers observe schooling but have only repeated noisy observations on cognitive skills, which affect productivity and are positively correlated with schooling, the effect of education on earnings declines with labour market experience, in line with our results. The human capital model is also consistent with our results if the earnings capacity invested in human capital during work declines with education. Mincer reports that returns to education decline with experience (see Willis, 1986, Table 10.5). Heckman et al, 2006, use US Census data and show that log earnings – experience profiles are parallel across schooling levels from 1940 to 1970 and converging from 1980 to 1990.

For this purpose, it is useful to briefly describe the labour market careers of firstborns and laterborns and to highlight the importance of labour mobility in the process of catching up of the latter with the former. Table 10 shows that firstborns are less mobile: they are 4.1 percentage points less likely than laterborns to have more than a single job in their careers, and more likely to be employed in their first job as white collar workers or as public sector employees.[21] These jobs are typically more stable than private sector jobs (see Clark and Postel-Vinay, 2009), and in some countries they are also associated to milder age earnings profiles.[22]

Define those who were still in their first job by age 30 as stayers and those who have moved to new jobs by that age as movers. Table 11 shows that firstborns at age 30 or 40 are more likely to be stayers.[23] For both firstborns and laterborns, the average log wage at age 30 can be written as $logW_{30} = p_{30} logW_{30}^{S} + (1-p) logW_{30}^{M}$, where the superscripts *S* and *M* are for stayers and movers, and *p* is the probability of being in the first job at age 30. In the case of firstborns, the log wage turns out to be equal to *9.506=0.421*9.300+0.579*9.656*. In the case of laterborns, it is equal to *9.513=0.356*9.250+0.644*9.659*. We notice that the average wage for laterborns at age 30 is only about 0.7% higher than the wage for firstborns (9.513 versus 9.506), in spite of the fact that firstborn stayers earn on average 5% more than laterborn stayers (9.3 versus 9.250). Since the average wage of movers is very similar across birth orders (9.656 versus 9.659), laterborns did catch up by age 30 because they were more likely to have moved by that age into better paid jobs: their probability of having done so was 0.644 rather than 0.579 for firstborns. Similar results hold for wages at age 40.[24] We conclude that firstborns start with a good match - sometimes a white collar or a public job - and stay in this match for a relatively long period. Laterborns instead struggle from initial low wages to higher wages by hopping quickly to new jobs.

Why do we observe these differences in labour market turnover? An important reason is education: since firstborns are better educated, they are more likely to locate a good initial match. An additional candidate factor, we believe, is that laterborns are more willing than firstborns to engage in risky behaviour and change employer more frequently. Allen et al, 2005, have shown that the relationship between turnover intentions and turnover is stronger for those lower in risk aversion. In support of

[21] Since 10.2 and 8.7 percent of laterborns are white collars and in the public sector respectively, the estimated percentage difference is equivalent to a 25.5 and a 35 percent gap. Our results are qualitatively unaltered when we add education as an additional control.

[22] Cappellari, 2002, for Italy and Hartog and Oosterbeek, 1993, for the Netherlands show that age earnings profiles are steeper in the private sector. Conversely, Dustmann and van Soest, 1998, show that profiles are steeper in the German public sector, and Disney et al., 2009 present mixed evidence for the UK. Following Zajonc, 1976, we speculate that firstborns may have had to share with parents the responsibility of growing younger siblings. This could have induced them to invest effort and parental networks to locate a good and stable first job and to keep it for a longer period of time. In support of this view, Table A3 in the Appendix shows that the probability that a firstborn lands a white collar or a public sector job as his first job increases with the number of siblings.

[23] Table A4 in the Appendix compares stayers and movers at age 30 and 40 and show that stayers – who are more likely to be firstborns – start on average their second job – if ever - at age 39, more than 17 years later than movers..

[24] We have also examined whether being firstborn has had any effect on experiencing unemployment but find no evidence that this is the case.

this view, the psychological literature has pointed out that laterborns tend to be more rebel and reckless with respect to firstborns,[25] who instead have a tendency to be more conscientious and self-disciplined (see Sulloway, 2007). Psychologists explain these differences by referring to the fact that while firstborns are endowed with higher parental resources,[26] laterborns are put under greater pressure to obtain the same returns from more limited resources and thus need to play riskier moves (see Wang et al, 2009).

To verify whether laterborns are less risk averse than firstborns, we use principal component analysis to extract the latent variable ϱ from the vector Γ, which includes five indicators of risk attitudes available in our data: whether the individual has ever bought private retirement accounts and life insurance packages, the body mass index and smoking and drinking habits.[27] Since this variable increases with risky health behaviours and decreases with the willingness to buy insurance and retirement accounts, we interpret it as a measure of risk taking. We regress ϱ on birth order and the other covariates and report our estimates in Table 12. We find that the effect of being firstborn on the willingness to take risks is negative and statistically significant, independently of whether we control or not for the mediating role of education.

We use these results to augment the human capital model so that it can account for findings a) and b).[28] The augmented Mincerian equation is given by

$$\ln w_{it} = a + b_1 S_i + c_1 x_{it} + b_2 S_i x_{it} + d_1 R_i + d_2 R_i x_{it} + f X_i + \lambda_j + \eta_i + \varepsilon_{it} \qquad (4)$$

where S, R and x are respectively years of schooling, risk taking attitudes and labour market experience. Dohmen et al, 2007, and Hartog et al, 2003, have shown that wages are increasing in risk taking attitudes $(d_1 > 0)$. We have shown that firstborns are more risk averse than laterborns, implying that if O_i is a dummy for being firstborn and $r_1 > 0$, then

$$R_i = r_0 - r_1 O_i + z_i \qquad (5)$$

[25] In his extensive monograph "Born to rebel", Sullloway, 1996, shows descriptive evidence that firstborns have always been more prone to support the status quo, and that laterborns have been more willing to challenge it. Nisbett, 1968, and Sulloway and Zweigenhaft, 2010, respectively show that laterborns are more likely to play risky sports than firstborns, and when playing the same sport they are more likely to carry out riskier moves. Zweigenhaft and von Ammon, 2000, show that being a laterborn positively affects the number of times a college student was arrested. Herrera et al., 2003, show how these findings mirror general beliefs about personality traits of first and laterborns.

[26] See also Lehmann et al., 2012, for evidence on differences in prenatal investments across first and laterborns.

[27] Smoking habits are captured by a dummy indicating whether the individual has ever smoked, and drinking habits by a dummy indicating whether the individual drinks alcohol on a daily basis.

[28] The learning model could also be augmented by positing that birth order captures other individual attitudes and non-cognitive skills accumulated before schooling (see Cunha and Heckman, 2007, and Heckman, Stixrud and Urzua, 2006). This extension would require, however, that employers can observe birth order. This seems unlikely in the presence of rules prohibiting discrimination.

Placing (5) into (4) yields

$$\ln w_{it} = (a + r_0 d_1) + b_1 S_i + (c_1 + r_0 d_2) x_{it} + b_2 S_i x_{it} - d_1 r_1 O_i - d_2 r_1 O_i x_{it} + f X_i + \lambda_j + \eta_i + \varepsilon_{it}$$

and by taking first differences we obtain

$$\Delta \ln w_{it} = (c_1 + r_0 d_2) + b_2 S_i - d_2 r_1 O_i + \Delta \varepsilon_{it} \qquad (6)$$

Using data from the Survey of Consumer Finances, Shaw, 1996, finds that wage growth is positively correlated with preferences for risk taking. In our setup, this implies that $d_2 > 0$, and that firstborns have lower earnings growth, in line with finding b). To explain finding a), notice that the entry wage is the wage at zero labour market experience (*t=0*), so that

$$\ln w_{i0} = (a + r_0 d_1) + b_1 S_i - d_1 r_1 O_i + f X_i + \lambda_j + \eta_i + \varepsilon_{i0} \qquad (7)$$

Furthermore, education is higher among firstborns, so the positive effect of being firstborn on the early wage requires that the positive effect of having higher education more than compensates the negative effect of being less willing to take risks.[29] The temporary nature of the advantage of being firstborn then follows from finding b).

6. Extensions and Robustness Checks

In this section, we provide a few extensions and sensitivities to the baseline results discussed in the previous section. First, we report in Table A5 the estimates of birth order effects when single children are excluded from the sample, and show that our results are hardly affected. Second, we investigate sources of heterogeneity in birth order effects by splitting the sample according to whether individuals lived in urban or rural areas at age ten, parental occupation at age ten was in blue or white collar jobs and finally between countries where the prevalent religion is protestant or catholic.

In rural areas, parental preferences for oldest children may have been stronger than in urban areas, with implications for labour market success. The estimates reported in Table A6 show that firstborns who were living in a rural area at age ten earn a higher premium in their first job with respect to laterborns than firstborns who lived in urban area. Yet, since the difference between the

[29] To see this, define $S_i = \pi_0 + \pi_1 O_i + \xi_i$ and substitute this in Eq.(7). We obtain that the marginal effect of being firstborn on earnings is $b_1 \pi_1 - d_1 r_1$.

estimated coefficients – reported in columns (1) and (2) of the table – is not statistically significant, we consider this evidence as suggestive at best.[30]

Next, we estimate our regressions separately for individuals who had parents in a blue collar or in a white collar job during childhood. We find that firstborns with a blue collar father earn a slightly higher premium over laterborns than firstborns with a white collar father. However, as in the previous case, we cannot reject the hypothesis that the estimated coefficients do not vary by parental group (Table A7). Finally, we report in Table A8 our estimates when the sample is separated in two groups of countries, depending on the prevailing religion in each country. Since protestants see success at work as a manifestation of the benevolence of God, we expect protestant parents to be less likely to favour first or later born children. Therefore, the wage premium in the initial job should be smaller in protestant than in catholic countries. Excluding Germany and Switzerland from the sample, because these two countries are not obviously protestant or catholic, we identify as protestant countries Denmark, Sweden and the Netherlands and as catholic countries the rest (Austria, Italy, Czechia, France and Spain). Our results do not confirm our priors, as we find that the effect of order of birth on wages does not significantly differ across groups of countries.

Conclusions

While there is substantial empirical research investigating the effects of birth order on educational attainment, little has been done to explore the effects on earnings. Some of the relatively few studies that have addressed this issue have found that the effects of birth order on earnings are rather negligible, in spite of the significant effects on educational attainment. This is puzzling if one believes that the key reason why birth order matters for wages is because it affects education.

We have used a sample of 4,280 European males born between 1935 and 1956 to study the effects of birth order on earnings over the life cycle. We have found that firstborns earn on average a 13.7% premium in their entry wage. This advantage, however, is completely gone by age 30. We have also found that being a firstborn has no statistically significant effect on earnings at age 50 or on current earnings, which are typically at a later age. We have estimated the effects of order of birth and education on earnings growth, measured at different points of the working life cycle, and found that both attract a negative and statistically significant coefficient, suggesting that education is not the only "mediator" of birth order effects on earnings.

We have interpreted these results by combining two facts: firstborns have both higher education and higher risk aversion than laterborns. Using these facts, for which we find support both in this paper and in the economic and psychological literature, we have argued that the observed patterns of earnings can be explained by differences in labour turnover. On the one hand, better education is a

[30] We test differences between coefficients by using the suest command in Stata 12.

key reason why firstborns start with a better match. On the other hand, the higher propensity to take risks explains why laterborns change jobs more frequently and enjoy higher wage growth than firstborns.

Our paper emphasizes the importance of using a life cycle approach in the study of the effects of birth order on earnings. This approach allows us to distinguish between temporary and permanent effects, unlike cross – sectional approaches that use a single observation of earnings for each individual. Since we have shown that the effect of birth order varies along the life cycle, choosing a single point in this cycle is likely to yield a misleading view of the relationship between birth order and earnings.

References

Allen, D.G., Weeks K.P. & Moffitt, K.R. (2005). Turnover intentions and voluntary turnover: the moderating roles of self-monitoring, locus of control, proactive personality, and risk aversion., *Journal of Applied Psychology,* 90, 980-990

Altonji, J. G., & Pierret, C. R. (2001). Employer learning and statistical discrimination. *The Quarterly Journal of Economics*, 116(1), 313-350.

Angelini, V., & Mierau, J. O. (2012). Childhood Health and the Business Cycle: Evidence from Western Europe. *HEDG working paper* n. 12/28, Department of Economics, University of York.

Bagger, J., Birchenall, J. A., Mansour, H., & Urzúa, S. (2013). Education, Birth Order, and Family Size , *NBER working paper* n. 19111.

Becker, G.S., & Lewis, H.G. (1973). Interaction between quantity and quality of children. *Journal of Political Economy*, 81 (2), pp. S279-S288.

Behrman, J. R., & Taubman, P. (1986). Birth Order, Schooling, and Earnings. *Journal of Labor Economics*, 4 (3), pp. S121-S145.

Björklund, A., & Jäntti, M. (2012). How important is family background for labor-economic outcomes?. *Labour Economics*, 19(4), 465-474.

Black, S.E., Devereux, P.J., & Salvanes, K.G. (2005). The more the merrier? The effect of family size and birth order on children's education. *Quarterly Journal of Economics,* 120 (2), pp. 669-700.

Brunello, G., Weber, G., & Weiss, C. T. (2012). Books are forever: Early life conditions, education and lifetime earnings in Europe, *IZA Discussion Paper* n. 6386.

Cappellari, L. (2002). Earnings dynamics and uncertainty in Italy: how do they differ between the private and public sectors?. *Labour Economics*, 9(4), 477-496.

Clark, A., & Postel-Vinay, F. (2009). Job security and job protection. *Oxford Economic Papers, 61(2)*, 207-239.

Cunha, F., & Heckman, J. J. (2007). The Technology of Skill Formation. *American Economic Review,* 97(2), 31-47.

De Haan, M. (2010). Birth order, family size and educational attainment, *Economics of Education Review*, 29(4), 576-588.

De Hann, M., Plug, E. & Rosero, J., (2012), Birth Order and Human Capital Development: Evidence from Ecuador, IZA Discussion Paper 6706

Dohmen, T. *et alia* (2007) Cross-sectional earnings risk and occupational sorting: The role of risk attitudes. *Labour Economics*, 14(6), 926-937.

Disney, R., Emmerson, C., & Tetlow, G. (2009). What is a Public Sector Pension Worth? *The Economic Journal*, 119(541), F517-F535.

Dustmann, C., & Van Soest, A. (1998). Public and private sector wages of male workers in Germany. *European Economic Review*, 42(8), 1417-1441.

Giuliano, P., & Spilimbergo, A. (2009). Growing up in a Recession: Beliefs and the Macroeconomy. *NBER working paper* n.15321.

Harmon, C., Oosterbeek, H. and Walker, I. (2003). The Returns to Education: Microeconomics. *Journal of Economic Surveys*, 17, 115–156.

Hartog, J., & Oosterbeek, H. (1993). Public and private sector wages in the Netherlands. *European Economic Review*, 37(1), 97-114.

Hartog, J., Plug, E, Diaz Serrano, L. and Vieria, J., (2003), Risk Compensation in Wage – a Replication, Empirical Economics, 28:639–647

Heckman, J. J., Lochner, L. J., & Todd, P. E. (2006). Earnings functions, rates of return and treatment effects: The Mincer equation and beyond. *Handbook of the Economics of Education*, 1, 307-458.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *NBER working paper* n. 12006.

Herrera, N. C., Zajonc, R. B., Wieczorkowska, G., & Cichomski, B. (2003). Beliefs about birth rank and their reflection in reality. *Journal of Personality and Social Psychology*, 85(1), 142.

Kantarevic, J., & Mechoulan, S. (2006). Birth Order, Educational Attainment, and Earnings An Investigation Using the PSID. *Journal of Human Resources*, 41(4), 755-777.

Kessler, D. (1991). Birth order, family size, and achievement: Family structure and wage determination. *Journal of Labor Economics,* 9 (4), pp. 413-426.

Lehmann, J. Y. K., Nuevo-Chiquero, A., & Vidal-Fernández, M. (2012).Explaining the birth order effect: The role of prenatal and early childhood investments. *IZA discussion paper* n. 6755.

Lindeboom, M., & Portrait, F. and Van den Berg, G. J., (2006). Economic conditions early in life and individual mortality. *The American Economic Review,* 96(1), 290-302.

Mantell, R. (2011) How birth order can affect your job salary. Available online at www.marketwatch.com/story/how-birth-order-can-affect-your-job-salary-2011-09-23

Modin, B. (2002). Birth order and mortality: a life-long follow-up of 14,200 boys and girls born in early 20th century Sweden. *Social science & medicine*, 54(7), 1051-1064.

Murphy, K. M., & Welch, F. (1990). Empirical age-earnings profiles. *Journal of Labor Economics*, 202-229.

Nisbett, R. E. (1968). Birth order and participation in dangerous sports. *Journal of personality and social psychology*, 8, 351.

Olneck, M. R., & Bills, D. B. (1979). Family configuration and achievement: Effects of birth order and family size in a sample of brothers. *Social Psychology Quarterly*, 135-148.

Oreopoulos, P., von Wachter, T., & Heisz, A. (2012). The short-and long-term career effects of graduating in a recession. *American Economic Journal: Applied Economics*, 4(1), 1-29.

Price, J. (2008). Parent-Child Quality Time Does Birth Order Matter?. *Journal of Human Resources*, 43(1), 240-265.

Shaw, K. L.,(1996). An Empirical Analysis of Risk Aversion and Income Growth. *Journal of Labor Economics,* 14 (4), 626-653.

Sulloway, F. J. (2007). Birth order and sibling competition. *The Oxford handbook of evolutionary psychology*, 297-311.

Sulloway, F. J. (1996). *Born to rebel: Birth Order, Family Dynamics, and Creative Lives*. Pantheon Books.

Sulloway, F.J., & Zweigenhaft, R.L. (2010). Birth Order and Risk Taking in Athletics: A Meta-analysis and Study of Major League Baseball Players. *Personality and Social Psychology Review*, 14, 402-416.

Wang, X. T., Kruger, D. J., & Wilke, A. (2009). Life history variables and risk-taking propensity. *Evolution and Human Behavior*, 30(2), 77-84.

Weiss, C. T.(2012). Two Measures of Lifetime Resources for Europe using SHARELIFE. *SHARE Working Paper Series*, 06-2012.

Willis, R. J. (1986). Wage determinants: A survey and reinterpretation of human capital earnings functions. In: O. Ashenfelter & R. Layard (eds.), *Handbook of Labor Economics*, volume 1, chapter 10, pages 525-602

Zajonc, R. B. (1976). Family configuration and intelligence: Variations in scholastic aptitude scores parallel trends in family size and the spacing of children. *Science*.

Zweigenhaft, R. L., & Von Ammon, J. (2000). Birth order and civil disobedience: A test of Sulloway's "born to rebel" hypothesis. *The Journal of Social Psychology*, 140(5), 624-627.

Tables

Table 1. Distribution of the number of brothers and sisters in the household at age ten

Number of siblings at age ten	Oldest child	Intermediate or youngest child
1	24.09	0
2	36.64	26.50
3	20.26	26.33
4	10.50	18.35
5	4.28	10.72
6	2.00	6.69
7	1.14	4.71
8	0.57	3.09
9	0.06	1.15
10+	0.46	2.45

Table 2. Summary statistics, by birth order

	Mean	Standard deviation	Number of obs.	Mean	Standard deviation	Number of obs.
		Oldest sibling			Other sibling	
First wage	11,786.46	11,931.81	1,752	10,577.37	13,055.06	2,528
Second wage	18,145.44	17,149.13	1,276	16,343.13	15,476.64	1,981
Third wage	22,307.85	19,460.54	849	20,541.04	16,688.16	1,364
Wage at 30	18,281.29	15,213.39	1,718	18,641.52	15,837.45	2,473
Wage at 40	22,162.24	17,644.81	1,741	21,586.36	16,490.78	2,503
Wage at 50	23,723.05	17,132.32	1,703	22,625.79	15,799.70	2,437
Current or last wage	23,546.84	15,161.50	1,752	22,787.13	15,169.60	2,528
Lifetime earnings net of pensions	8,844.11	5,580.00	1,752	8,676.27	5,486.53	2,528
Not employed at age 30	0.019	0.138	1,752	0.022	0.146	2,528
Not employed at age 40	0.006	0.079	1,752	0.010	0.099	2,528
Not employed at age 50	0.028	0.165	1,752	0.036	0.186	2,528
Age when first job started	19.602	4.165	1,752	18.587	4.060	2,528
Age when last job ended	58.163	4.429	1,752	57.786	4.404	2,528
Oldest child	1	-	1,752	0	-	2,528
Only child	0.241	0.428	1,752	0	-	2,528
Number of siblings (including the interviewed person)	2.512	1.456	1,752	3.917	1.944	2,528
Mother in the house at ten	0.965	0.183	1,752	0.972	0.165	2,528
Father in the house at ten	0.913	0.282	1,752	0.930	0.255	2,528
Foster mother in the house at ten	0.021	0.142	1,752	0.011	0.105	2,528
Foster father in the house at ten	0.032	0.176	1,752	0.017	0.128	2,528
Grandparents in the house at ten	0.147	0.354	1,752	0.106	0.308	2,528
Other relatives in the house at ten	0.059	0.236	1,752	0.049	0.215	2,528
Other non-relatives in the house at ten	0.016	0.125	1,752	0.022	0.146	2,528
Hunger episodes before age 15	0.031	0.174	1,752	0.042	0.200	2,528
Parents smoke, drank or had mental problems	0.691	0.462	1,752	0.700	0.458	2,528
At least one parent died before turning 35	0.038	0.192	1,752	0.017	0.129	2,528
Breadwinner at ten is blue collar	0.661	0.474	1,752	0.722	0.448	2,528
Lived in rural area	0.378	0.485	1,752	0.439	0.496	2,528
Years of education	12.593	4.091	1,752	11.487	4.250	2,528
Age of mother at birth	23.876	3.960	764	27.671	4.707	692

Source. SHARE survey waves 1, 2 and 3.

Table 3. Birth order effects on education, by number of siblings. Dependent variable: number of years of schooling

	Two siblings	Three siblings	Four siblings	All siblings
Oldest child	0.749***	0.676***	0.706**	0.645***
	(0.205)	(0.248)	(0.336)	(0.123)
Number of siblings	-	-	-	-0.123***
				(0.032)
Observations	1,312	1,021	648	4,280
R-squared	0.243	0.253	0.320	0.254

Notes: All regressions include dummies for: cohort, country, mother in the house at 10, father in the house at 10, foster mother in the house at 10, foster father in the house at 10, grandparents in the house at 10, other relatives in the house at 10, hunger episodes by age 15, parents smoked, drank or had mental problems, at least one parent died by age 35, breadwinner occupation at age 10, lived in rural area at age 10. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table 4. Birth order effects on real earnings. By family size and pooling sizes. Dependent variable: log real wage

	Entry wage 2 siblings	Entry wage 3 siblings	Entry wage 4 siblings	Entry wage all siblings	Wage in current or last job 2 siblings	Wage in current or last job 3 siblings	Wage in current or last job 4 siblings	Wage in current or last job all siblings
Oldest child	0.135**	0.145**	0.186*	0.137***	0.004	-0.023	0.030	-0.011
	(0.056)	(0.064)	(0.098)	(0.033)	(0.031)	(0.039)	(0.057)	(0.020)
Number of siblings	-	-	-	-0.032***	-	-	-	-0.021***
				(0.009)				(0.005)
Observations	1,312	1,021	648	4,280	1,312	1,021	648	4,280
R Squared	0.236	0.255	0.250	0.233	0.244	0.264	0.227	0.210

Notes. All regressions include dummies for: cohort, country, mother in the house at 10, father in the house at 10, foster mother in the house at 10, foster father in the house at 10, grandparents in the house at 10, other relatives in the house at 10, hunger episodes by age 15, parents smoked, drank or had mental problems, at least one parent died by age 35, breadwinner occupation at age 10, lived in rural area at age 10. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table 5. Birth order effects on earnings – by pooling family sizes

	Entry wage	Wage at 30	Wage at 40	Wage at 50	Wage in current or last job	Lifetime earnings
Oldest child	0.137***	-0.024	-0.002	0.007	-0.011	0.000
	(0.033)	(0.026)	(0.023)	(0.022)	(0.020)	(0.019)
Number of siblings	-0.032***	-0.011	-0.015**	-0.018***	-0.021***	-0.015***
	(0.009)	(0.007)	(0.007)	(0.006)	(0.005)	(0.006)
Observations	4,280	4,191	4,244	4,140	4,280	4,280
R-squared	0.233	0.221	0.208	0.205	0.210	0.266

Notes: All regressions include dummies for: cohort, country, mother in the house at 10, father in the house at 10, foster mother in the house at 10, foster father in the house at 10, grandparents in the house at 10, other relatives in the house at 10, hunger episodes by age 15, parents smoked, drank or had mental problems, at least one parent died by age 35, breadwinner occupation at age 10, lived in rural area at age 10. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table 6. Birth order effects on earnings – with and without imputation

	Entry wage no imputation	Entry wage with imputation	Wage in current or last job no imputation	Wage in current or last job with imputation
Oldest child	0.129***	0.137***	-0.011	-0.011
	(0.040)	(0.033)	(0.020)	(0.020)
Number of siblings	-0.034***	-0.032***	-0.021***	-0.021***
	(0.011)	(0.009)	(0.005)	(0.005)
Observations	3,262	4,280	4,278	4,280
R-squared	0.247	0.233	0.211	0.210

Notes: All regressions include dummies for: cohort, country, mother in the house at 10, father in the house at 10, foster mother in the house at 10, foster father in the house at 10, grandparents in the house at 10, other relatives in the house at 10, hunger episodes by age 15, parents smoked, drank or had mental problems, at least one parent died by age 35, breadwinner occupation at age 10, lived in rural area at age 10. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table 7. Birth order effects on earnings – with and without controls for age of mother at birth

	Entry wage	Entry wage	Wage in current or last job	Wage in current or last job
Oldest child	0.187***	0.175***	0.005	-0.015
	(0.060)	(0.057)	(0.035)	(0.033)
Number of siblings	-0.047***	-0.048***	-0.022**	-0.022**
	(0.017)	(0.017)	(0.010)	(0.010)
Age of mother at birth	0.003	-	0.005	-
	(0.007)		(0.004)	
Observations	1,456	1,456	1,456	1,456
R-squared	0.262	0.261	0.204	0.202

Notes: All regressions include dummies for: cohort, country, mother in the house at 10, father in the house at 10, foster mother in the house at 10, foster father in the house at 10, grandparents in the house at 10, other relatives in the house at 10, hunger episodes by age 15, parents smoked, drank or had mental problems, at least one parent died by age 35, breadwinner occupation at age 10, lived in rural area at age 10. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table 8. Birth order effects on wage growth over the life cycle

	Wage at 30 – Entry wage	Wage at 40 – Entry wage	Wage at 50 – Entry wage	Wage in current or last job - first wage
Oldest child	-0.162***	-0.135***	-0.142***	-0.148***
	(0.033)	(0.035)	(0.035)	(0.036)
Number of siblings	0.022**	0.016	0.013	0.011
	(0.010)	(0.010)	(0.010)	(0.010)
Observations	4,191	4,244	4,140	4,280
R-squared	0.082	0.096	0.106	0.137

Notes: All regressions include dummies for: cohort, country, mother in the house at 10, father in the house at 10, foster mother in the house at 10, foster father in the house at 10, grandparents in the house at 10, other relatives in the house at 10, hunger episodes by age 15, parents smoked, drank or had mental problems, at least one parent died by age 35, breadwinner occupation at age 10, lived in rural area at age 10. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table 9. Birth order and education effects on wage growth over the life cycle

	Wage at 30 – Entry wage	Wage at 40 – Entry wage	Wage at 50 – Entry wage	Wage in current or last job - first wage
Oldest child	-0.131***	-0.103***	-0.110***	-0.119***
	(0.033)	(0.035)	(0.035)	(0.036)
Years of schooling	-0.049***	-0.050***	-0.048***	-0.045***
	(0.004)	(0.004)	(0.004)	(0.005)
Number of siblings	0.016*	0.010	0.007	0.005
	(0.010)	(0.010)	(0.010)	(0.010)
Observations	4,191	4,244	4,140	4,280
R-squared	0.110	0.123	0.132	0.156

Notes: All regressions include dummies for: cohort, country, mother in the house at 10, father in the house at 10, foster mother in the house at 10, foster father in the house at 10, grandparents in the house at 10, other relatives in the house at 10, hunger episodes by age 15, parents smoked, drank or had mental problems, at least one parent died by age 35, breadwinner occupation at age 10, lived in rural area at age 10. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table 10. Birth order, number of jobs held and type of first job

	Had more than one job	First job was full time	First job was white collar	First job was in public sector
Oldest child	-0.041***	-0.004	0.029***	0.036***
	(0.014)	(0.006)	(0.010)	(0.010)
Number of siblings	0.001	0.002	-0.003	-0.000
	(0.004)	(0.002)	(0.002)	(0.003)
Observations	4,280	4,280	4,275	4,280
R-squared	0.0654	0.030	0.118	0.062

Notes: All regressions include dummies for: cohort, country, mother in the house at 10, father in the house at 10, foster mother in the house at 10, foster father in the house at 10, grandparents in the house at 10, other relatives in the house at 10, hunger episodes by age 15, parents smoked, drank or had mental problems, at least one parent died by age 35, breadwinner occupation at age 10, lived in rural area at age 10. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table 11. Birth order effects on the probability of being still in the first job at selected ages

	Still in first job at 30	Still in first job at 40
Oldest child	0.058***	0. 051***
	(0.016)	(0.015)
Number of siblings	-0.004	-0.003
	(0.004)	(0.004)
Observations	4,191	4,244
R-squared	0.084	0.083

Notes: All regressions include dummies for: cohort, country, mother in the house at 10, father in the house at 10, foster mother in the house at 10, foster father in the house at 10, grandparents in the house at 10, other relatives in the house at 10, hunger episodes by age 15, parents smoked, drank or had mental problems, at least one parent died by age 35, breadwinner occupation at age 10, lived in rural area at age 10. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table 12. Birth order, education and the propensity to take risks

Oldest child	-0.071**	-0.058**
	(0.028)	(0.028)
Years of schooling	-	-0.021***
		(0.003)
Number of siblings	0.006	0.004
	(0.007)	(0.007)
Observations	3,929	3,929
R-squared	0.189	0.197

Notes: All regressions include dummies for: cohort, country, mother in the house at 10, father in the house at 10, foster mother in the house at 10, foster father in the house at 10, grandparents in the house at 10, other relatives in the house at 10, hunger episodes by age 15, parents smoked, drank or had mental problems, at least one parent died by age 35, breadwinner occupation at age 10, lived in rural area at age 10. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Appendix

Table A1. Summary statistics

	Mean	Standard deviation	Number of observations
First wage	11,072.31	12,619.95	4,280
Second wage	17,049.22	16,173.88	3,257
Third wage	21,218.86	17,819.29	2,213
Wage at 30	18,493.85	15,583.83	4,191
Wage at 40	21,822.60	16,974.00	4,244
Wage at 50	23,077.15	16,367.90	4,140
Current or last wage	23,098.11	15,169.12	4,280
Lifetime earnings net of pensions	8,744.98	5,524.95	4,280
Not employed at age 30	0.021	0.143	4,280
Not employed at age 40	0.008	0.091	4,280
Not employed at age 50	0.033	0.178	4,280
Age when first job started	19.002	4.133	4,280
Age when last job ended	57.940	4.418	4,280
Oldest child	0.409	0.492	4,280
Only child	0.099	0.298	4,280
Number of siblings (including the interviewed person)	3.342	1.891	4,280
Mother in the house at ten	0.969	0.173	4,280
Father in the house at ten	0.923	0.266	4,280
Foster mother in the house at ten	0.015	0.121	4,280
Foster father in the house at ten	0.023	0.150	4,280
Grandparents in the house at ten	0.123	0.328	4,280
Other relatives in the house at ten	0.053	0.224	4,280
Other non-relatives in the house at ten	0.019	0.138	4,280
Hunger episodes before age 15	0.037	0.190	4,280
Parents smoke, drank or had mental problems	0.696	0.460	4,280
At least one parent died before turning 35	0.026	0.158	4,280
Breadwinner at ten is blue collar	0.697	0.460	4,280
Lived in rural area	0.414	0.493	4,280
Years of education	11.940	4.220	4,280
Age of mother at birth	25.679	4.727	1,456

Source: SHARE survey waves 1, 2 and 3.

Table A2. Birth order effects on the probability of not being employed at different ages

	Not employed at 20	Not employed at 25	Not employed at 30	Not employed at 35	Not employed at 40	Not employed at 50
Oldest Child	0.072***	0.029***	-0.004	-0.000	-0.002	-0.004
	(0.015)	(0.010)	(0.005)	(0.003)	(0.003)	(0.006)
Number of Siblings	-0.008*	-0.000	-0.001	0.000	0.000	0.002
	(0.004)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)
Observations	4,280	4,280	4,280	4,280	4,280	4,280
R-squared	0.119	0.069	0.021	0.014	0.018	0.023

Notes: All regressions include dummies for: cohort, country, mother in the house at 10, father in the house at 10, foster mother in the house at 10, foster father in the house at 10, grandparents in the house at 10, other relatives in the house at 10, hunger episodes by age 15, parents smoked, drank or had mental problems, at least one parent died by age 35, breadwinner occupation at age 10, lived in rural area at age 10. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table A3. Birth order, number of jobs held and type of first job. By number of siblings

	First job white collar, at most 3 siblings	First job was full time, 4 siblings or more	First job was in public sector, at most 3 siblings	First job was in public sector, 4 siblings or more
Oldest child	0.019	0.040**	0.031**	0.049**
	(0.013)	(0.019)	(0.013)	(0.021)
Number of siblings	-0.008	-0.005	-0.004	-0.002
	(0.009)	(0.004)	(0.009)	(0.004)
Observations	2,755	1,520	2,755	1,525
R-squared	0.147	0.159	0.076	0.083

Notes: All regressions include dummies for: cohort, country, mother in the house at 10, father in the house at 10, foster mother in the house at 10, foster father in the house at 10, grandparents in the house at 10, other relatives in the house at 10, hunger episodes by age 15, parents smoked, drank or had mental problems, at least one parent died by age 35, breadwinner occupation at age 10, lived in rural area at age 10. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table A4. Labour turnover and earnings growth

Age 30		Age 40	
	Mean		Mean
Still in first job		*Still in first job*	
Log wage at 30	9.273	Log wage at 40	9.395
Log first wage	9.058	Log first wage	9.000
Log second job	9.857	Log second wage	9.776
Log third job	9.944	Log third wage	9.850
Age started first job	20.589	Age started first job	20.462
Age ended first job	51.214	Age ended first job	56.130
Age started second job	39.225	Age started second job	47.666
Age started third job	43.883	Age started third job	50.975
Not in first job anymore		*Not in first job anymore*	
Log wage at 30	9.658	Log wage at 35	9.869
Log first wage	8.604	Log first wage	8.698
Log second job	9.232	Log second wage	9.319
Log third job	9.622	Log third wage	9.655
Age started first job	17.865	Age started first job	18.421
Age ended first job	21.471	Age ended first job	23.545
Age started second job	21.977	Age started second job	22.842
Age started third job	27.784	Age started third job	29.023

Table A5. Birth order effects, excluding single children

	Entry wage	Wage in current or last job	Wage in current or last job – entry wage
Oldest Child	0.140***	-0.015	-0.155***
	(0.035)	(0.021)	(0.038)
Number of Siblings	-0.034***	-0.021***	0.013
	(0.010)	(0.006)	(0.011)
Observations	3,858	3,858	3,858
R-squared	0.225	0.205	0.135

Notes: All regressions include dummies for: cohort, country, mother in the house at 10, father in the house at 10, foster mother in the house at 10, foster father in the house at 10, grandparents in the house at 10, other relatives in the house at 10, hunger episodes by age 15, parents smoked, drank or had mental problems, at least one parent died by age 35, breadwinner occupation at age 10, lived in rural area at age 10. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table A6. Birth order effects on earnings, by rural and urban areas

	Entry wage. Rural	Entry wage. Urban	Wage in current or last job. Rural	Wage in current or last job. Urban	Wage in current or last job – entry wage. Rural	Current or last wage - first wage. Urban
Oldest child	0.168***	0.121***	-0.021	-0.010	-0.189***	-0.130***
	(0.051)	(0.044)	(0.029)	(0.027)	(0.056)	(0.048)
Number siblings	-0.015	-0.048***	-0.011	-0.031***	0.004	0.016
	(0.014)	(0.013)	(0.008)	(0.008)	(0.015)	(0.014)
Observations	1,772	2,508	1,772	2,508	1,772	2,508
R-squared	0.212	0.264	0.271	0.182	0.130	0.161

Notes: All regressions include dummies for: cohort, country, mother in the house at 10, father in the house at 10, foster mother in the house at 10, foster father in the house at 10, grandparents in the house at 10, other relatives in the house at 10, hunger episodes by age 15, parents smoked, drank or had mental problems, at least one parent died by age 35, breadwinner occupation at age 10, lived in rural area at age 10. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table A7. Birth order effects on earnings, by parental occupation

	Entry wage. White collar breadwinner	Entry wage. Blue collar breadwinner	Wage in current or last job. White collar breadwinner	Wage in current or last job. Blue collar breadwinner	Wage in current or last job – entry wage. White collar breadwinner	Current or last wage - first wage. Blue collar breadwinner
Oldest child	0.121**	0.144***	-0.038	0.004	-0.159**	-0.140***
	(0.058)	(0.041)	(0.035)	(0.024)	(0.063)	(0.044)
Number siblings	-0.028	-0.034***	-0.030***	-0.018**	-0.002	0.016
	(0.019)	(0.011)	(0.011)	(0.006)	(0.021)	(0.012)
Observations	1,297	2,983	1,297	2,983	1,297	2,983
R-squared	0.224	0.212	0.190	0.201	0.137	0.139

Notes: All regressions include dummies for: cohort, country, mother in the house at 10, father in the house at 10, foster mother in the house at 10, foster father in the house at 10, grandparents in the house at 10, other relatives in the house at 10, hunger episodes by age 15, parents smoked, drank or had mental problems, at least one parent died by age 35, breadwinner occupation at age 10, lived in rural area at age 10. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table A8. Birth order effects on earnings, by prevailing religion in the country

	Entry wage. Catholic	Entry wage. Protestant	Wage in current or last job. Catholic	Wage in current or last job. Protestant	Wage in current or last job – entry wage. Catholic	Current or last wage - first wage. Protestant
Oldest child	0.139**	0.134**	-0.044	0.014	-0.183***	-0.120*
	(0.047)	(0.060)	(0.030)	(0.030)	(0.052)	(0.062)
Number siblings	-0.020	-0.050***	-0.019***	-0.013	0.001	0.037**
	(0.013)	(0.016)	(0.007)	(0.009)	(0.014)	(0.017)
Observations	2,320	1,318	2,320	1,318	2,320	1,318
R-squared	0.215	0.212	0.153	0.085	0.14	0.152

Notes: All regressions include dummies for: cohort, country, mother in the house at 10, father in the house at 10, foster mother in the house at 10, foster father in the house at 10, grandparents in the house at 10, other relatives in the house at 10, hunger episodes by age 15, parents smoked, drank or had mental problems, at least one parent died by age 35, breadwinner occupation at age 10, lived in rural area at age 10. Robust standard errors in parentheses. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Chapter 4

Hungry Today, Happy Tomorrow?
Childhood Conditions and Self-Reported Wellbeing Later in Life[1*]

by

Marco Bertoni

(University of Padova and CEP, LSE)

Abstract

We use anchoring vignettes to show that, on data for eleven European countries, exposure to episodes of hunger in childhood leads people to adopt lower subjective reference points to evaluate satisfaction with life in adulthood. This is consistent with the satisfaction treadmill theory of hedonic adaptation, and highlights that failure to consider reporting heterogeneity will result in downward-biased estimates of the negative effects of starvation in childhood on the levels of wellbeing later in life. These findings underline the importance of considering issues of interpersonal comparability when studying the determinants of subjective wellbeing.

Keywords: subjective wellbeing, childhood conditions, anchoring vignettes.
JEL Codes: C42, I31, J13.

* [1] I am extremely grateful for supervision from Giorgio Brunello, for advice from Danilo Cavapozzi, and for suggestions from Viola Angelini, Luca Corazzini, Marta de Philippis, Andrea Moro, Omar Paccagnella, Enrico Rettore, Elisabetta Trevisan and Guglielmo Weber, and participants at seminars in Padua and Venice. All errors are my own. This paper uses data from SHARELIFE release 1, as of November 24th 2010 and SHARE release 2.5.0, as of May 24th 2011. 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, 211909 and SHARE-LEAP, 227822).

Introduction

This paper studies the effects of childhood conditions on self-reported wellbeing later in life. There is wide evidence in economics showing that events taking place early in life affect personality traits, education, late-life health, socio-economic conditions and wellbeing[2]. Understanding the long reach of events happening in critical periods of human development (Cuhna and Heckman, 2007) is thus of uttermost importance to identify proximate causes of successful lives (Layard et al., 2013). Our study is focused in particular on the experience of hunger episodes in childhood, as a case of extreme deprivation, and our outcome of interest is subjective wellbeing. Policy makers and economists are devoting an increasing amount of effort to the analysis of self-reported happiness and to subjective indicators of life satisfaction, with the aim of considering a definition of welfare that goes beyond strictly economic measures[3]. Yet, as underlined by Clark et al., 2005, the first question to be asked when studying what makes a happy person concerns the meaning of "being happy": a crucial issue of interpersonal comparability limits in fact the use of subjective measures to inform policy decisions. In psychometrics, this problem is known as differential item functioning (Holland and Wainer, 1993), and refers to the fact that people interpret and use the same reporting scale differently.

The aim of the paper is twofold. First, we want to assess the effect of early-life starvation on the reference points used by individuals to rate life satisfaction. As highlighted by Deaton, 2008, evaluation of subjective wellbeing is a relative one. People compare their situation with a subjective benchmark, a "shifting standard" that depends on comparison with peers and with one's past experiences. According to the satisfaction treadmill theory of hedonic adaptation (Kahneman, 1999, and Frederick, 2007), the experience of extreme deprivation may lead people to develop lower aspirations regarding the level of life achievements to consider as satisfying. As a consequence, when asked to evaluate a given situation on a discrete and bounded rating scale (e.g. a scale going from 1 to 5), individuals exposed to hunger episodes will assign it a better judgement than people not exposed to hunger, even if that situation represents the same level of happiness for both groups. To the best of our knowledge, there is no empirical evidence documenting whether there is an effect of childhood condition on individual reporting scales, and this paper contributes to fill this gap.

---

[2] For instance, Ichino and Winter-Ebmer, 2004, Akbulut-Yuksel, 2009, and Kesternich et al., 2012, find negative effects of exposure to WWII on education, health and labour market outcomes. Meng and Qian, 2009, Lindeboom et al., 2010, Havari and Peracchi, 2011, Neelsen and Stratmann, 2011, Pinger et al., 2011, and Kesternich et al., 2013, study the effects of exposure to childhood hunger on several outcomes later in life, including longevity, education, health and the share of budget devoted to food purchases, while Lindeboom et al., 2006, van den Berg et al., 2010, and Angelini and Mierau, 2013, relate exposure to negative macroeconomic conditions at birth to longevity, health and reaction to adverse shocks later in life. Finally, both Frijters et al., 2011, and Layard et al., 2013, analyse life-cycle models of subjective wellbeing, that link childhood conditions to happiness later in life.

[3] See Layard, 2005, 2006 and 2013, for thorough discussions. The United Nation's World Happiness Report series, and the newly realized OECD Better Life Index are just some policy efforts in this sense.

On top of the importance of this question to qualify how hedonic adaptation to negative life events works, modelling individual-specific response scales allows us to evaluate the effects of hunger episodes on wellbeing levels without worrying about interpersonal comparability of self-reported happiness evaluations: if people exposed to hunger adopt lower reference points, estimates that do not take reporting style differences into consideration will be biased towards finding positive effects.

Commonly used methods to deal with scale bias when longitudinal data are available include the use of individual fixed effects models, that help dealing with time-invariant reporting heterogeneity. Still, results from Angelini et al., 2011, highlight that the scale people use to rate their conditions is a time-varying one, and these techniques are not of much help to deal with cross-sectional data. We instead rely on a vignettes approach. In an anchoring vignettes questionnaire, respondents first evaluate their own situation in a given domain, then they rate a series of descriptions of situations of hypothetical persons, the vignettes, using the same rating scale applied for the self-assessment. We can use answers to vignette questions as an anchor to properly model individual reporting scales and, in turn, to filter subjective evaluations from reporting heterogeneity[4].

We use data on eleven European countries from the Survey of Health, Ageing and Retirement in Europe, SHARE, a longitudinal and multidisciplinary survey of the European population aged 50+. There are two main advantages of using SHARE for this analysis: the second wave of the survey contains a vignette questionnaire on life satisfaction, that can be used to model reporting heterogeneity, while the third wave collects retrospective information on people's life histories, including data on specific periods of hunger[5] and other childhood experiences, and on family background.

Our main empirical result is that experience of hunger in childhood leads people to shift their subjective reference points downwards, i.e., to give a higher rating to the same latent level of wellbeing. We also find a long run scarring effect of starvation in childhood on the level of happiness later in life, but contrarily to the extant literature we are the first to derive our results from a model that takes differences in reporting styles into account. Comparing our estimates with the ones from a model that does not allow for individual-specific evaluation scales, we conclude that failure to properly treat reporting heterogeneity leads to underestimate the effects of negative childhood conditions on late-life wellbeing.

---

4 This approach was introduced in social sciences by King et al., 2004, in the context of political efficacy. Kapteyn et al., 2007, and Angelini et al., 2011, are illustrative examples concerning disability conditions, while Bago d'Uva, 2008, Kok et al., 2012, and Peracchi and Rossetti, 2012, study self-reported health and depression. In the context of life satisfaction, Angelini et al., 2012 and 2013, look at satisfaction with life in general, and Bonsang and Van Soest, 2012a and 2012b, look at satisfaction with social contacts, job and income.

5 These measures have been validated against potential issues of recall bias (see Garrouste and Paccagnella, 2011, and Havari and Mazzonna, 2011). However, given the self-reported nature of this information, if more pessimistic people report suffering of hunger more easily then the effects of hunger on reporting scales that we are estimating are lower bounds to the true ones.

Beyond contributing to the economic and psychological literature on hedonic adaptation, showing an example of satisfaction treadmill, our findings are also relevant for the economic literature that wants to model the determinants of life satisfaction over the life cycle, as we highlight the need to consider the effects of childhood conditions on individual-specific reporting scales to draw meaningful conclusions on what predict a satisfactory life using subjective data.

The paper unfolds as follows. Section 1 presents the data we use and some descriptive statistics. Our econometric model is described in Section 2, and Section 3 illustrates our main empirical results. We discuss the validity of the vignettes approach for our analysis in Section 4, while Section 5 presents some sensitivities and extensions to our main results. Conclusions follow thereafter.

1. The Data

This paper exploits the Survey of Health, Ageing and Retirement in Europe, SHARE, that collects detailed longitudinal information on household composition, socio-economic status, health and wellbeing of the population aged 50+ of several European countries.

On the one hand, we draw information on adulthood conditions, on self-reported wellbeing and on anchoring vignettes from the COMPARE subsample of the second wave of SHARE, that was collected between 2006 and 2007. On the other hand, we rely on the information on childhood environment contained in SHARELIFE, the third wave of the survey, that was carried out between 2008 and 2009. Our final sample is composed of 4,950 individuals who took part in both the COMPARE and the SHARELIFE surveys, were born between 1920 and 1956 and were residing in Belgium, the Czech Republic, Denmark, France, Germany, Greece, Italy, the Netherlands, Poland, Spain and Sweden at the time of the wave 2 interview[6]. Unfortunately, about 30% of the initial COMPARE sample is lost in SHARELIFE because of panel attrition. To test for endogenous attrition[7], in Table A1 in the Appendix we compare for each country the mean values of several variables[8] from wave 2 computed in the full sample and in the selected sample that we consider. The mean values of most variables look very similar across the two samples, and we cannot detect any selection pattern that is common for all countries. We conclude that endogenous attrition is not an issue for the data at hand.

Data on life satisfaction and anchoring vignettes were collected in SHARE within the COMPARE project. After completion of the main interview, in eleven countries a representative subset of

[6] We also drop a small minority of individuals with missing values on any of the variables included in our analysis. Information on life satisfaction, anchoring vignettes, and hunger episodes is missing for less than 1% of the sample.
[7] Bonsang and Van Soest, 2012a, show that the COMPARE sample is comparable to the full SHARE sample of wave 2.
[8] These include gender, age, log annual household income, being affected by one or more of a list of limitations in the Activities of Daily Living (ADL), having an ISCED 4 or higher educational qualification, being in a couple, being employed, being retired, the self-assessment and anchoring vignettes evaluation of satisfaction with life.

respondents were asked to complete on their own a paper-and-pencil questionnaire containing self-assessment questions on satisfaction with life and on several other health and disability domains. For each domain, brief descriptions of the conditions of hypothetical persons were also included (anchoring vignettes), and individuals were asked to rate these as well. Beyond the self-assessment question, two vignette questions on life satisfaction were also present (see Angelini et al. 2012 and 2013), and all life satisfaction ratings had to be provided on the same five-points ordinal scale. The exact wording of the three questions is reported in Figure 1, together with the reporting scale that respondents were asked to use, while Figure 2 presents the distribution of answer to self-assessment and vignette questions on life satisfaction in our sample. There is substantial variation in vignette evaluations across individuals in our sample, suggesting that reporting heterogeneity cannot be easily ignored, and the global ordering of vignette evaluations suggests that Carrie's conditions are generally evaluated as describing a higher level of life satisfaction than John's.

Other adulthood variables we consider include demographic information, educational levels (primary, secondary or post-secondary qualifications), dummy variables for being in a couple, for suffering of one or more of a list of limitations in Activities of Daily Living (ADL) or in Instrumental Activities of Daily Living (IADL), for being employed and for retirement. We also use information on annual household income and on household net wealth[9]. Descriptive statistics for these variables are shown in Table 1.

We derive information on childhood from the third wave of SHARE, SHARELIFE, that collects retrospective data on the entire lives of respondents, including early life conditions and family, employment, housing, and health histories. SHARELIFE has the advantage of providing data on life histories that are comparable across several European countries, but there is also a potential threat of recall bias, that is common in retrospective studies (see Smith, 2009). Nonetheless, validation studies carried out by Garrouste and Paccagnella, 2011, and Havari and Mazzonna, 2011, show that the state-of-the-art elicitation methods used in SHARELIFE, based on life history calendars, greatly reduced the incidence of recall bias. Amongst early life conditions, our main interest lies in the experience of hunger episodes in childhood, i.e. from birth until age 15. The SHARLIFE questionnaire explicitly asks each individual to recall whether she experienced "a period when you suffered from hunger", and if so to indicate the exact years when the hunger episode started and stopped. In previous studies, Havari and Peracchi, 2011, Pinger et al., 2011, Kesternich et al., 2012, Halmdienst and Winter-Ebmer, 2013, and Attanasio et al., 2014, validated the reliability of this self-reported measure against historical events. Figure 3 shows the distribution of people in a hunger

[9] Financial variables are measured in PPP terms at German prices of 2006, and divided by household size. As in Angelini et al., 2013, we transform financial variables using the arcsinh(.) function, allowing for zeroes and for negative values that show up for wealth.

episode that started in childhood in our sample, by country and year[10]. Hunger episodes are most commonly reported during World War II years, from 1939 until 1945, but further patterns related to historical events can be traced out in the data. For instance, we see that among Spanish people hunger episodes start to be reported already from the mid-1930s, in coincidence with the Spanish Civil War. We also observe peaks in hunger that correspond to the Greek famine of 1941-1942, the Dutch famine of 1944 and the German famine of 1945-1948 (see Pinger et al., 2011, for further details)[11].

We also exploit SHARELIFE to derive information on other conditions experienced by respondents in their childhood. First, following Mazzonna, 2011, we use principal component analysis to extract a single indicator of childhood socio-economic status from a vector of commonly used proxies (see Brunello, Weber and Weiss, 2012) that comprises the following variables: the number of books at home at age 10, occupation of the main breadwinner in the household at age 10, the number of rooms per person and the presence of an inside toilet and of running water in the accommodation where the individual was living at age 10. Besides hunger episodes, like Bohacek and Myck, 2010, and Halmdienst and Winter-Ebmer, 2013, we consider a set of other relevant childhood events that might be related to starvation and to wellbeing later in life. These include having lived in an orphanage or with foster parents, being relocated during childhood because of war, being dispossessed of the family's house, land, business or of other properties during childhood[12]. Like Kesternich et al., 2012, we also consider an indicator for exposure to war events in childhood, that varies by cohort and country[13], but we are not able to distinguish between broad exposure to war and living in specific combat areas, as they instead do. Finally, we develop indicators for living in a rural area at age 10[14], for whether the father or the mother were absent from the household at age 10, for having "troubled" parents, that smoked, drank or had mental health problems, and for having any siblings while in childhood. Descriptive statistics for these variables are also reported in Table 1, while Table 2 shows marginal effects of a Probit model that

[10] This figure differs quantitatively from the ones reported in other studies mentioned above, as we consider only hunger episodes that started before age 15, but the qualitative picture is very similar.

[11] Kesternich et al., 2013, further validate this self-reported measure by looking at post-WWII Germany and showing that the regional and temporal variation in self-reported hunger episodes closely mirrors differences in centralised food supply rations between German regions during that period of time.

[12] Unlike Bohacek and Myck, 2010, we cannot consider also whether the individual was a victim of prosecution, as in the SHARELIFE questionnaire there is no indication on the timing of prosecution and we would not be able to distinguish between instances of prosecutions that happened in childhood and later on.

[13] Like Kesternich et al., 2012, we consider as exposed to war in childhood all individuals born until 1945 in all non-neutral countries (neutral countries were Sweden and Denmark) except for Germany, where war ended in 1948, and who were younger than 15 when the war started, i.e. in 1939 in all countries except for Spain, where we the civil war started in 1936.

[14] Neelsen and Stratmann, 2011, examine the Greek famine of 1941-42, and show that hunger incidence was more severe in urban than rural areas.

relates this set of childhood conditions to the occurrence of hunger[15]. Results from this descriptive analysis confirm that most of the childhood conditions considered are significantly linked to starvation in infancy, and that hunger is also, but not only, the result of poverty. We are going to include these variables as controls in all further analysis.

2. Empirical Methods

The aim of our empirical analysis is to exploit the availability of anchoring vignette questions to estimate the effect of childhood hunger on the subjective scales individuals use to report wellbeing. This will allow us to filter subjective evaluations from differential item functioning, and to evaluate the effects of hunger on the levels of adulthood wellbeing without issues of reporting heterogeneity. To this end, we use the Heterogeneous Thresholds (or Hierarchical) Ordered Probit model, Hopit, introduced by King et al., 2004. We first describe the econometric model, then we discuss the conditions under which our findings can be given a causal interpretation.

The vignettes approach to identification in presence of reporting heterogeneity relies on two assumptions. The first one, response consistency, posits that individuals use the same response scale to evaluate their own condition and the ones described in vignette questions. The second assumption, vignette equivalence, states that there are no differences across respondents in the perception of the level of life satisfaction described by each vignette (Bago d'Uva et al., 2011), so that differences in vignette evaluations only reflect differences in reporting styles.

As highlighted by Peracchi and Rossetti, 2013, differential item functioning "is essentially a problem of identification in ordered response models, where the observed responses are derived from latent continuous random variables, discretized through a set of *heterogeneous* cut-off points". Under vignette equivalence, variation in responses to anchoring questions allows identification of individual-specific cut-off points used to report wellbeing on a discrete scale. In turn, response consistency allows to use this set of individual-specific thresholds to model the self-assessments free from issues of interpersonal comparability. The validity of the two assumptions has been widely debated in the literature, with mixed findings. We will come back to this point while discussing our empirical findings.

We model the latent level of life satisfaction assessed by individual *i*, Y_i^*, as follows:

$$\begin{aligned} Y_i^* &= \beta X_i + \varepsilon_i \\ \varepsilon_i &\approx N(0,1) \end{aligned} \qquad (1)$$

[15] Controls for cohort and country of birth are included as well, and their effects mirror the evidence presented in Figure 2.

In (1), X_i is a vector of individual observable variables that includes exposure to hunger and the other childhood covariates described in the previous section, while ε_i is an error term that is independent of X_i and follows a standard normal distribution[16]. In the data we do not observe Y_i^*, as individuals are asked to report their life satisfaction on a five points ordinal scale. The following rule is assumed to relate the observed life satisfaction level Y_i to Y_i^*:

$$Y_i = j \; if \; \xi_i^{j-1} < Y_i^* \leq \xi_i^j, j = 1, \ldots, 5 \qquad (2)$$

To model subjective response scales, the Hopit model allows thresholds ξ_i to depend on the same set of individual covariates included in (2), according to the following specification:

$$\begin{aligned} &\xi_i^0 = -\infty; \; \xi_i^5 = \infty; \\ &\xi_i^1 = \gamma^1 X_i + \eta_i; \\ &\xi_i^j = \xi_i^{j-1} + \exp\left(\gamma^j X_i\right), j=2, \; 3, \; 4 \\ &\eta_i \approx N\left(0, \sigma_\eta^2\right) \end{aligned} \qquad (3)$$

In (3), the exponential specification is only needed to grant monotonicity of the individual-specific thresholds, while as in Kapteyn et al., 2007, we extend the basic Hopit model to allow for an individual-specific random effect in the threshold equations, η_i, that allows for unobserved correlation between the evaluations of the three life satisfaction assessments.

The parameters in β and γ^1 would not be identified using self-assessments questions alone, while the parameters in $\gamma^j, j = 2, 3, 4$ would be identified only via the exponential functional form restriction.

In the Hopit model identification exploits the fact that each individual is also asked to rate two vignette questions on the same measurement scale used for the self-assessment. Under vignette equivalence, the latent evaluation of the k-th vignette, Z_{ki}^*, does not systematically vary over individuals i, so that

$$\begin{aligned} &Z_{ki}^* = \theta_k + \nu_{ki}, \\ &\nu_{ki} \approx N\left(0, \sigma_\nu^2\right) \end{aligned} \qquad (5)$$

Vignettes are evaluated on the same five points ordinal scale used for self assessment, and the observed rating of vignette k, Z_{ki}, is linked to Z_{ki}^* as follows:

$$Z_{ki} = j \;\; \text{if} \;\; \xi_i^{j-1} < Z_{ki}^* \leq \xi_i^j, \; \text{j} \; = \; 1, \ldots, 5 \qquad (6)$$

[16] Location and scale normalization are achieved by setting the constant term to 0 and the variance of the error term ε_i to 1.

Under response consistency, the same thresholds apply to both self-assessment and vignette questions. This assumption allows to link the two components of the model: vignette evaluations are used to identify the parameters in θ and γ, that fully describe the set of discretizing cut-offs ξ_i. Given ξ_i, the vector β is identified through the self-reports. The full model is estimated via maximum likelihood.

Causal interpretation of the effects of childhood hunger on the levels and on the scale used to report life satisfaction in adulthood deserves further discussion, beyond the validity of vignette equivalence and response consistency. While episodes of starvation in childhood temporally pre-date wellbeing evaluation later in life, so that no issue of reverse causation arises, causal interpretation of our findings might be jeopardized by the presence of other confounding factors. As discussed in Section 1, experience of childhood hunger is at least partly supply-driven, as it is associated to the exogenous occurrence of famines and central rationing of food supply. On the other hand, as shown in Table 2, hunger is also driven by exposure to World War II and to other childhood events, and the incidence of starvation is higher among individuals coming from a low socio-economic background. As a consequence, any unconditional relation between hunger and self-reported wellbeing can hardly be interpreted as causal. The strategy we exploit to deal with endogeneity is simple selection-on-observables[17]: we assume that conditioning on the wealth of childhood covariates described in Section 1 mops up the remaining unwanted correlation between any other observable and unobservable determinant of both late-life wellbeing and childhood hunger that is not included in the model, so that the remaining correlation between hunger and self-reported wellbeing can be interpreted in a causal sense. Of course, unless this untestable assumption holds our findings can only be interpreted in a descriptive sense, as conditional correlations.

3. Main Empirical Results

The empirical results described in this section are based on the random effects Hopit model presented in Section 2. We also compare these findings with a baseline model that does not allow for interpersonal variation in reporting scales, akin to an Ordered Probit model, to assess the relevance of reporting heterogeneity. Table 3 shows selected estimation results for exposure to hunger, while full outcomes are presented in Table A2 in the Appendix. In both tables, results for the baseline model are shown in column 1, while column 2 reports parameter estimates for the self-assessment equation in the Hopit model, and columns 3, 4, 5 and 6 report estimates for the

[17] Other papers have exploited the sources of exogenous variation described in Section 1, famines and wars, as instruments for childhood hunger (e.g. see Havari and Peracchi, 2011, and Pinger et al., 2011). However, no IV-Hopit model - allowing instrumental identification of the effect of hunger on both the levels and the scale used to report subjective wellbeing - is available so far. We leave the development of such models for further research, and choose to make do with selection-on-observables for this paper.

parameters in the equation for threshold 1, 2, 3 and 4, respectively. A formal likelihood ratio test strongly rejects the baseline model vis-à-vis the heterogeneous threshold one (p-value = 0.000), implying that the presence of heterogeneity in reporting styles is statistically relevant.

Our key finding is that exposure to hunger significantly affects the scale used to report life satisfaction[18]. Looking column 6 in Table 3, we see that there is a negative shift in the upper discretizing threshold for people who suffered of starvation early in life, i.e., they use the highest point on the ordinal scale to rate the same level of satisfaction with life that would receive a lower evaluation from people not exposed to hunger[19]. We interpret this result in the light of the satisfaction treadmill theory of hedonic adaptation (see Kahneman, 1999, and Frederick, 2007). This theory posits that, since no absolute scale to evaluate wellbeing exists, people carry out subjective evaluations with respect to individual-specific reference points, that are assumed to depend on the conditions of close peers and on one's past experiences. Therefore, individuals who experienced extreme deprivation in childhood may evaluate the same life achievements more positively, as they could have developed lower aspirations for what having a satisfactory life means.

This positive rescaling effect does not mean that people experiencing hunger are more satisfied with their lives. Indeed, we find that the opposite is true: childhood hunger is negatively linked to happiness levels later in life, and the relation is strongly significant. We estimate a long-run negative scarring effect of childhood hunger on satisfaction with life in adulthood, that is consistent with other studies on negative childhood conditions and late-life wellbeing (e.g. see Havari and Peracchi, 2011, and Kesternich et al., 2012, both using data from SHARELIFE), but is derived in a more general framework that takes reporting heterogeneity into account. Comparing results from the baseline and the Hopit models, we see that failure to consider reporting heterogeneity will result in downward biased estimates of the effect of childhood hunger on the levels of life satisfaction, highlighting that subjective measures of wellbeing may only partially reflect the effects of life events on satisfaction with life in presence of rescaling effects.

To quantitatively assess the relevance of these rescaling effects, we turn to a counterfactual simulation. Results are shown in Figure 4. The upper graph plots the distribution of life satisfaction that is predicted by our model. We derive this figure by first computing for each respondent the predicted values of latent life satisfaction, $\hat{Y}_i^*$, and of her specific cut-off points, $\hat{\xi}_i$, given her individual characteristics, and then by plotting the resulting distribution of life satisfaction measured on the five points ordinal scale. The middle and lower graphs, instead, show the counterfactual

[18] The p-value of a test for joint significance of hunger exposure in the four threshold equations is reported at the bottom of Table 4: the null hypothesis is strongly rejected in the data.

[19] Since the hunger indicator we are using is self-reported, this effect is a lower bounds to the true effect on reporting scales if more pessimistic people are more prone to report hunger. As a consequence, even our corrections for reporting heterogeneity on the levels of subjective wellbeing are lower bounds to the true ones.

distributions of life satisfaction that we would observe if we assigned to each individual the cut-off points she would have used had she, or had she not, experienced childhood hunger, respectively, while leaving all her other baseline characteristics and her self-assessment equation unaltered. According to our estimates, the share of the sample reporting to be "very satisfied" with life increases by a non-negligible 12% using the reporting scales that would hold under exposure to hunger with respect to the complementary case, while the exact same share of people would report to be "neither satisfied, nor dissatisfied"[20].

Looking at the full set of estimated parameters, reported in Table A2 in the Appendix, it is reassuring to see that the same patterns described for hunger hold for socio-economic status in childhood as well: kids that start from a disadvantaged background are less satisfied with life later on, but tend to rate the same situations more positively. In our view, this finding strengthens our interpretation in terms of a satisfaction treadmill[21]. Furthermore, significant differences on wellbeing levels and on reporting styles are found across people from different countries, and the overall patterns are comparable to the ones described by Angelini et al., 2013. Finally, the parameter associated with John's vignette is more negative than the one associated with Carrie's, consistently with the global ordering of the two vignettes, while the standard deviation of the individual random effects is statistically significant, bringing support to the extended Hopit model vis-à-vis its simple formulation[22].

4. Discussing the Identifying Assumptions

Before showing some sensitivities and extensions to our main results, it is worth discussing the crucial issue of validity of response consistency and vignette equivalence. Response consistency states that individuals adopt the same subjective scale to evaluate satisfaction with their own life and the situations described in each vignette. Several tests for this assumption were proposed in the literature. Kapteyn et al., 2011, test this hypothesis by conducting a survey experiment that relies on longitudinal data on health conditions and self-reported health status. They construct individual-specific vignettes illustrating the past situation of each respondent, deriving the relevant information from her answers to questions posed in previous waves of the panel, and ask her to rate the vignette. Comparing current vignette evaluations with past self-assessments, they find mixed support for response consistency, depending on the domain of interest. A growing number of studies suggests

[20] On the other hand, it is not surprising to see that the top and bottom graph are very similar, as only a small fraction of the sample was exposed to childhood hunger.
[21] Other childhood conditions have limited impact on both reporting scales and life satisfaction, once socio-economic status and hunger episodes are taken into account.
[22] This finding is also confirmed by a formal likelihood ratio test, that rejects the baseline Hopit model in favour of the extended one (p-value = 0.000).

instead to validate response consistency using objective measures of the concept of interest. For instance, Van Soest et al., 2011, consider self-perceived drinking problems, and show that people's evaluation of their own drinking problems and of vignettes describing a drinking level equal to their own one are strongly aligned, supporting response consistency[23]. Angelini et al., 2013, however, highlight that such tests can hardly be implemented when it comes to evaluating satisfaction with life in general, as this is a multidimensional concept that cannot be unquestionably measured using a single objective indicator. They instead show that, in the SHARE data, people whose situation in each of the domains described in a vignette is analogous to the situation depicted in the vignette give equal evaluations of satisfaction with their own life and of that of the person described in the vignette, bringing strong support to response consistency for the case of satisfaction with life.

Even when response consistency holds, the validity of the vignettes approach requires a further assumption, vignette equivalence. To test vignette equivalence one needs to show that the level of life satisfaction of the person described in a vignette is perceived equally by each respondent, irrespectively of his or her personal characteristics. Murray et al., 2003, highlight that a minimal requirement for vignette equivalence to hold is that individuals systematically order the vignettes in the same way[24]. Kristensen and Johansson, 2008, further suggest that, in a cross-country study, vignette equivalence is unlikely to hold if we find that results from models estimated pooling all countries and separating groups of countries that share different sets of values and social norms are very different from one another[25]. Instead, finding similar results should be supportive of vignette equivalence. When more than one vignette is available, Bago d'Uva et al., 2011, propose to use one vignette to anchor the assessment of the other ones, and to test whether personal characteristics used to model the thresholds enter significantly in the evaluation of the other available vignettes. Under the hypothesis of response consistency, evidence in this sense would suggest violations of vignette equivalence. Using data on mobility limitations and cognitive functions, they do not find strong support for vignette equivalence. Peracchi and Rossetti, 2013, propose a statistical test for the joint validity of the two assumptions that exploits the fact that the Hopit model is over-identified if both response consistency and vignette equivalence hold. Using data on several health domains, they find mixed support for the validity of the two assumptions, depending on the domain and sample considered. Still, both tests are based on the further assumptions of correct model specification and no omitted variables, i.e., they may reject the null hypothesis for reasons that differ from failure of

[23] A similar test was proposed by Datta Gupta et al., 2010, and Bago d'Uva et al., 2011, studying health, cognitive functions and mobility limitations. However, their results do not unambiguously point in favour of response consistency.

[24] Still, Bago d'Uva et al., 2011, highlight that interpersonal differences in the levels of vignette evaluations are not inconsistent with equal ranking of the vignettes.

[25] While this is especially true for the estimated vignette levels (θ_1 and θ_2) and for the coefficients associated with the country dummies, this conclusion is less likely to hold for the effects of other individual-level covariates, as different effects across the split samples may just reflect heterogeneous effects by country.

the two identifying conditions. Thus, as no formal test of vignette equivalence is available yet, as in Angelini et al., 2013, we produce informal evidence on the validity of vignette equivalence in our data.

Following the ordering test proposed by Murray et al., 2003, in our data we see that only 8% of respondents rate vignettes inconsistently with the global ordering[26], which is reassuring. This finding holds also across groups of countries that share similar values, as identified by the Inglehart-Wezel values map (see Kristensen and Johansson, 2008). As in Angelini et al., 2013, we consider three groups of countries: ex-communist countries (Czech Republic and Poland), catholic countries (Belgium, France, Italy and Spain) and protestant countries (Denmark, Germany, the Netherlands and Sweden)[27]. We replicate our estimation outcomes dropping individuals who rate vignettes inconsistently with the global ordering, and results (not shown) are fully comparable with the baseline, suggesting that our findings are not driven by violations of vignette equivalence that would result in inconsistent vignette ordering. We also re-estimate our model separately for the three groups of countries identified according to the Inglehart-Wezel values map and described above, leaving Germany as the baseline country in every group. Table A3 in the Appendix shows that estimated vignette levels and country dummies from the pooled and split samples are in line with each other in terms of both ordering and magnitude, confirming findings by Angelini et al., 2013, and bringing support to vignette equivalence.

5. Sensitivities and Extensions

This section proposes some sensitivity checks and some extensions to our main empirical results. First, it is interesting to understand whether our results on the effects of hunger on self-reported wellbeing are mediated by adulthood conditions that depend on hunger episodes and are commonly known to determine satisfaction with life. Evidence from SHARE presented by Havari and Peracchi, 2011, and Pinger et al., 2011, shows that childhood hunger is related to lower educational achievement and worse health and labour market outcomes in adulthood. In turn, using the SHARE data Angelini at al., 2013, show that these characteristics are linked with both the levels and the scale used to report life satisfaction. We thus introduce a set of adulthood outcomes as mediating variables in our model[28], and evaluate by how much the effects of childhood hunger change with

[26] That is, they rate Carrie's vignette as representing a lower satisfaction level than John's (see Figure 2).
[27] Since it is an orthodox country, we drop Greece for this analysis.
[28] The variables we consider are taken from the second wave of SHARE and were described in Section 1. Although Angrist and Pischke, 2008, refer to mediators as "bad controls", a similar strategy is exploited, for instance, by Giuliano and Spilimbergo, 2013, to assess the effects of growing up in a recession on beliefs in adulthood, by Frjiters et al., 2011, and Layard et al., 2013, in the context of modelling life-cycle dynamics of subjective wellbeing, and by Halmdienst and Winter-Ebmer, 2013, to seek for mediators of the effects of early-life shocks on health later in life, using SHARELIFE data.

respect to our baseline. Results for hunger are reported in Table 4, while complete estimation outcomes are shown in Table A4 in the Appendix. We see that the effects of hunger on reporting scales and wellbeing levels survive even when we include the mediating variables[29], suggesting that their role is not a central one.

Next, we consider potential issues of dynamic selection. Looking at the Dutch potato famine of 1846-1847, Lindeboom et al., 2010, show that nutritional conditions in infancy negatively affect mortality age. If more optimistic people survive longer after an episode of hardship, or if survivors have characteristics that make them more prone to report life satisfaction using more positive scales, sample selection may bias our results towards finding positive reporting effects. To understand whether survival is linked to reporting styles and life satisfaction levels, we estimate a Probit regression of the probability of deceasing between SHARE wave 2, conducted in 2006, and the successive wave, conducted in 2008, on vignette evaluation, self-reported satisfaction with life and a set of controls measured at baseline, in 2006[30]. Results in Table 5 do not suggest that reporting styles and life satisfaction levels are correlated with death probabilities. Although these results are based on surviving in old age, and not on surviving to hunger episodes in childhood, they are still reassuring about concerns of dynamic selection on the basis of reporting styles and life satisfaction levels. Furthermore, in Table 6 we drop from our sample individuals born before 1930 or before 1935, amongst whom dynamic selection should be more serious[31], and results are fully comparables with our baseline.

We provide further sensitivities in Table A5 in the Appendix. Panel A excludes from our sample people who migrated from country of birth, a result that may be linked to the experience of hunger, and results are still in line with the ones presented in Table 3. Finally, in Panel B and Panel C we respectively drop people residing in neutral countries during World War II (i.e. Sweden and Denmark) and people born after 1949, as episodes of hunger were less common within these groups. Again, the patterns detected are comparable with our baseline.

We next explore potential sources of heterogeneity in the effects of hunger on wellbeing, taking into account the effects of hunger on reporting scales[32].

[29] Reassuringly, the effects of the mediating variables are in line with results presented by Angelini et al., 2013. Results of counterfactual simulations from this model, available from the author, are also similar the ones presented in Figure 4.
[30] The sample considered excludes wave 2 respondent for whom no information is available at wave 3 because of unit non-response. We checked that unit non-response in wave 3 does not depend on life satisfaction and vignette evaluation in separate regressions (not shown). Similar results (not shown) are obtained when we consider survival until wave 4, conducted in 2011. In that case, however, we had to remove Greece, that did not take part in the fourth wave of the survey.
[31] If recall bias is also stronger for older individuals, results from the subsample of younger individuals should be less subject to this issue as well.
[32] Our results are mainly descriptive, and this evidence should be viewed as suggestive.

We first ask whether there are gender differences in the effects of childhood hunger on wellbeing. Evidence coming mainly from developing countries[33] shows that the impacts of negative childhood shocks are larger for girls than for boys, and that this is partly explained by differential compensation mechanisms carried out by parents, who tend to favour survival of sons vs. daughters. Table 7 reports results when we split the sample by gender, and we see that scarring effects on wellbeing levels are larger for girls, and smaller and only marginally significant for boys.

According to Graham and Oswald, 2010, the dynamics of adaptation to life events depend on hedonic capital, i.e., the stock of psychological resources available to an individual. In their theoretical model, hedonic capital depends on a collection of stock-like variables affecting individual wellbeing, like interpersonal relationships, psychological traits, and social status, that help to smooth wellbeing when a shock hits. In this context, the presence of siblings may contribute to the generation of hedonic capital: sharing the same negative experiences and encouraging each other, children with siblings should be able to cope with starvation better than single children. In turn, higher income may help to buffer the effect of hunger episodes. We test these two hypothesis by splitting the sample between single children and individuals with siblings, and between respondents with an index of socio-economic status below and above the median level in the sample. Results presented in Table 8 show that scarring effects on wellbeing levels are indeed stronger for only children. Finally, Table 9 confirms that people from higher socio-economic background suffer less from hunger episodes in childhood, consistently with the idea of a buffering effect of income.

Conclusions

Current attention in public policy and economics is devoted to developing a deeper understanding of the factors promoting happiness throughout life. Using data from eleven European countries, our paper looks at the effects of childhood hunger on self-reported wellbeing later in life. Using anchoring vignettes, we are able to disentangle the effects of early life starvation on the levels and on the subjective scale used to report wellbeing in adulthood.

We contribute to the literature by showing that people exposed to hunger in childhood have lower subjective reference points against which they evaluate life satisfaction, as predicted by the satisfaction treadmill theory of hedonic adaptation. This is a new finding, that sheds light on the determinants of endogenous benchmarks used to evaluate subjective wellbeing, on which little is known so far. Consequently, using Hopit models we are able to provide the first evidence of negative long-run scarring effects of childhood hunger on the levels of wellbeing that is free from issues of interpersonal comparability, and we show that failure to consider reporting heterogeneity

[33] For instance, see Maccini and Yang, 2009, for a review.

will result in estimates that are biased downwards. Hence, our results raise awareness on the importance of considering differences in reporting styles when studying the long reach of childhood conditions onto late-life wellbeing.

We are aware that the validity of the conclusions from this paper strongly depends on the identifying assumptions behind the empirical analysis. Research that provides more robust evidence on the validity of the vignettes approach will help to validate our findings, as well as research developing methods to exploit instrumental variables techniques within the vignettes framework.

References

Akbulut-Yuksel, M., 2009. *Children of war: The long-run effects of large-scale physical destruction and warfare on children*. IZA Discussion Paper 4407.

Angelini, V., Cavapozzi, D. and Paccagnella, O., 2011. Dynamics of reporting work disability in Europe. *Journal of the Royal Statistical Society: Series A,* 174(3), 621-638.

Angelini, V., Cavapozzi, D., Corazzini, L. and Paccagnella, O., 2012. Age, Health and Life Satisfaction Among Older Europeans. *Social Indicators Research*, 105(2), 293-308.

Angelini, V., Cavapozzi, D., Corazzini, L. and Paccagnella, O., 2013. Do Danes and Italians Rate Life Satisfaction in the Same Way? Using Vignettes to Correct for Individual-Specific Scale Biases. *Oxford Bulletin of Economics and Statistics*, forthcoming.

Angelini, V. and Mierau, J.O., 2012. *Childhood Health and the Business Cycle: Evidence from Western Europe.* HEDG Working Papers n.12/28.

Angrist, J, D. and Pischke, J. S., 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Attanasio O., Brugiavini A., Trevisan E. and Weber G., 2014. The consequences of financial hardship (and recessions) on income and welfare. In: A. Brugiavini and G. Weber (eds.): *Longer-term consequences of the Great Recession on the lives of Europeans*, Oxford University Press: Oxford (forthcoming)

Bago d'Uva, T., Lindeboom, M., O'Donnell, O. and van Doorslaer, E., 2008. Does reporting heterogeneity bias the measurement of health disparities?. *Health economics*, 17(3), 351-375.

Bago d'Uva, T., Lindeboom, M., O'Donnell, O. and van Doorslaer, E., 2011. Slipping Anchor?: Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity. *Journal of Human Resources*, 46(4), 875-906.

Bohacek, R. and Myck, M., 2011. *Long shadows of history: persecution in central Europe and its labor market consequences*. IZA Discussion Paper 6130.

Bonsang, E. and van Soest, A., 2012a. Satisfaction with Social Contacts of Older Europeans. *Social Indicators Research*, 105(2), 273-292.

Bonsang, E. and van Soest, A., 2012b. Satisfaction with job and income among older individuals across European countries. *Social Indicators Research*, 105(2), 227-254.

Brunello, G., Weber, G., and Weiss, C. T., 2012. *Books are forever: Early life conditions, education and lifetime earnings in Europe*. IZA Discussion Paper n. 6386.

Clark, A., Etilé, F., Postel-Vinay, F., Senik, C. and van der Straeten, K., 2005. Heterogeneity in Reported Well Being: Evidence from Twelve European Countries. *The Economic Journal*, 115(502), C118-C132.

Cunha, F. and Heckman, J. J., 2007. The Technology of Skill Formation. *The American Economic Review*, 2007, 97(2), 31-47.

Datta Gupta, N., Kristensen, N. and Pozzoli, D., 2010. External Validation of the Use of Vignettes in Cross-Country Health Studies. *Economic Modelling*, 2010, 27 (4), 854-865.

Deaton, A., 2008. Income, health and wellbeing around the world: Evidence from the Gallup World Poll. *The Journal of Economic Perspectives*, 22(2), 53-72.

Frederick, S., 2007. *Hedonic Treadmill*. H-Baumeister (Encyc)-45348.qxd. 419-420.

Frijters, P., Johnston, D. W. and Shields, M. A., 2011. *Destined for (Un)Happiness: Does Childhood Predict Adult Life Satisfaction?* IZA Discussion Paper 5819.

Garrouste, C. and Paccagnella, O., 2012. *Data Quality: Three Examples of Consistency Across SHARE and SHARELIFE Data*. SHARELIFE Methodology.

Giuliano, P., and Spilimbergo, A., 2013. Growing up in a recession. *Review of Economic Studies*, forthcoming.

Graham, L. and Oswald, A. J., 2010. Hedonic capital, adaptation and resilience. *Journal of Economic Behavior and Organization*, 2010, 76(2), 372-384.

Halmdienst, N. and Winter-Ebmer, R., 2013. *Long-Run Effects of Childhood Shocks on Health in Late Adulthood*. NRN working papers, 2013-01.

Havari, E., and Mazzonna, F., 2011. *Can we trust older people's statements on their childhood circumstances? Evidence from SHARELIFE*. SHARE Working Paper Series 05-2011

Havari, E. and Peracchi, F., 2011. *Childhood circumstances and adult outcomes: Evidence from SHARELIFE*. EIEF Working Paper 1115.

Holland, P.W. and Wainer, H., 1993. *Differential Item Functioning*. Hillsdale (NJ): Erlbaum.

Ichino A., and Winter-Ebmer R., 2004. The long-run educational cost of World War Two. *Journal of Labor Economics*, 22, 57-86

Kahneman, D., 1999. Objective happiness. In Kahneman, D., Diener, E. and Schwartz, N., *Well-being: The foundations of hedonic psychology* (3–25). New York: Russell Sage

Kapteyn, A., Smith, J.P. and van Soest, A., 2007. Vignettes and Self-Reports of Work Disability in the United States and the Netherlands. *The American Economic Review*, 97(1), 461-473.

Kapteyn, A., Smith, J. P. and van Soest, A., 2011. *Anchoring Vignettes and Response Consistency*, RAND Working Papers 840.

Kesternich, I., Siflinger, B., Smith, J. P. and Winter, J. K., 2012. The effects of World War II on economic and health outcomes across Europe. *Review of Economics and Statistic*s, forthcoming.

Kesternich, I., Siflinger, B., Smith, J. P. and Winter, J. K., 2013. *Individual Behavior as a Pathway Between Early-Life Shocks and Adult Health. Evidence from Hunger Episodes in Post-War Germany*. Rand Working Papers 1015

King, G., Murray, C. J. L., Salomon, J. A. and Tandon, A., 2004. Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research. *American Political Science Review*, 98, 191–207.

Kok, R., Avendano, M., Bago d'Uva, T. and Mackenbach, J., 2012. Can reporting heterogeneity explain differences in depressive symptoms across Europe?. *Social indicators research*, 105(2), 191-210.

Kristensen, N., and Johansson, E., 2008. New evidence on cross-country differences in job satisfaction using anchoring vignettes. *Labour Economics*, 15(1), 96-117.

Layard, R., 2005. *Happiness: Lessons from a New Science*. London: Penguin.

Layard, R., 2006. Happiness and Public Policy: a Challenge to the Profession. *The Economic Journal*, 116(510), C24-C33.

Layard, R., 2013. Mental health: the new frontier for labour economics *IZA Journal of Labor Policy*, 2(2).

Layard, R., Clark, A., Cornaglia, F. and Powdthavee, N., 2013. *What Predicts a Successful Life? A Life-Course Model of Well-Being*, IZA Discussion Paper 7682.

Lindeboom, M., Portrait, F., and van den Berg, G. J., 2006. Economic Conditions Early in Life and Individual Mortality. *The American Economic Review*, 96(1), 290-302.

Lindeboom, M., Portrait, F., and van den Berg, G. J., 2010. Long-run effects on longevity of a nutritional shock early in life: The Dutch Potato famine of 1846-1847, *Journal of Health Economics*, vol. 29(5), 617-629-

Maccini, S. and Yang, D,. 2009. Under the Weather: Health, Schooling, and Economic Consequences of Early-Life Rainfall. *The American Economic Review*, 99(3), 1006-1026.

Mazzonna, F., 2011. *The long-lasting effects of family background: a European cross-country comparison*. Munich Centre for the Economics of Aging Working Paper Nr. 245-2011.

Meng, X. and Qian, N., 2009. *The Long Term Consequences of Famine on Survivors: Evidence from a Unique Natural Experiment using China's Great Famine.* NBER Working Papers 14917.

Murray, C. J. L., Ozaltin, E., Tandon, A., Salomon, J.A., Sadana, R. and Chatterji, S., 2003. Empirical evaluation of the anchoring vignettes approach in health surveys. In Murray, C.J.L. and Evans, D.B, *Health systems performance assessment: debates, methods and empiricism* (369-399). Geneva: World Health Organisation.

Neelsen S., and Stratmann T., 2011. Effects of prenatal and early life malnutrition: Evidence from the Greek famine, *Journal of Health Economics*, 30, 479-488.

Peracchi, F. and Rossetti, C., 2012. Heterogeneity in health responses and anchoring vignettes. *Empirical Economics*, 42(2), 513-538.

Peracchi, F. and Rossetti, C., 2013. The heterogeneous thresholds ordered response model: Identification and inference. *Journal of the Royal Statistical Society: Series A,* forthcoming.

Pinger, P., Schoch, J. and van den Berg, G.J., 2011. *Instrumental variable estimation of the causal effect of hunger early in life on health later in life.* IZA Discussion Paper 6110

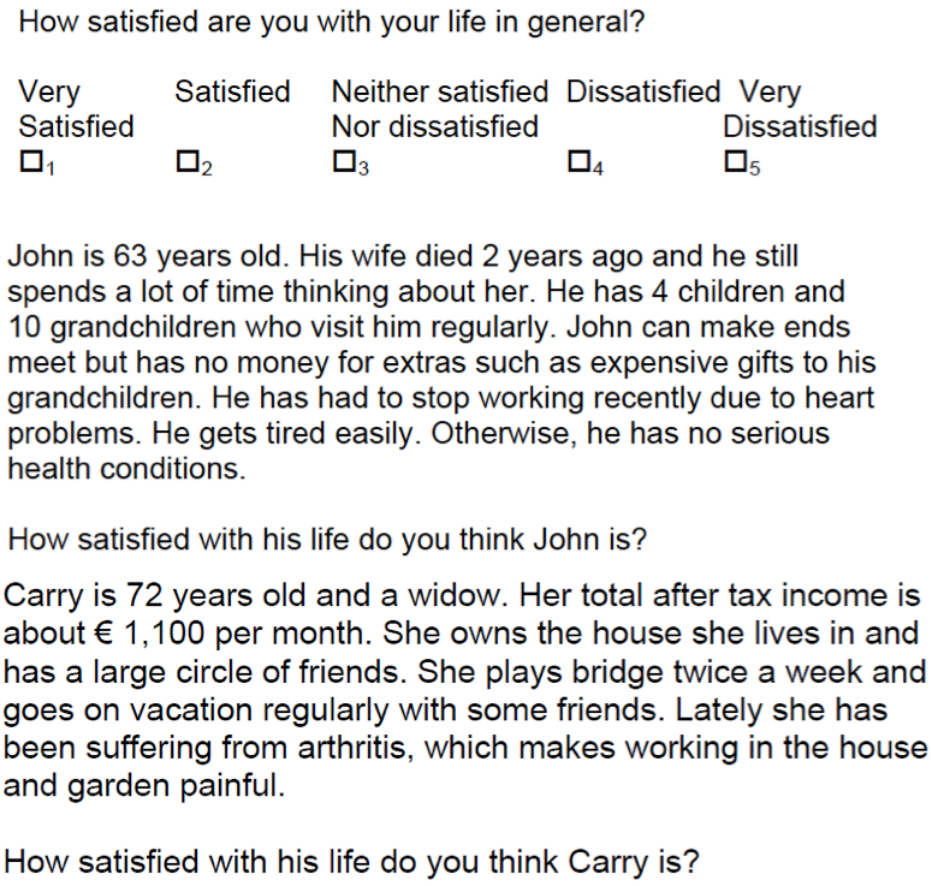
Smith, J. P, 2009. Reconstructing childhood health histories. *Demography*, 46(2), 387-403.

Van den Berg, G. J., Deeg, D., Lindeboom, M. and Portrait, F., 2010. The Role of Early-Life Conditions in the Cognitive Decline due to Adverse Events Later in Life, *The Economic Journal*, 120(548), F411-F428.

Van Soest, A., Delaney, L., Harmon, C., Kapteyn, A. and Smith, J. P., 2011. Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions. *Journal of the Royal Statistical Society Series A*, 174(3), 575-595.

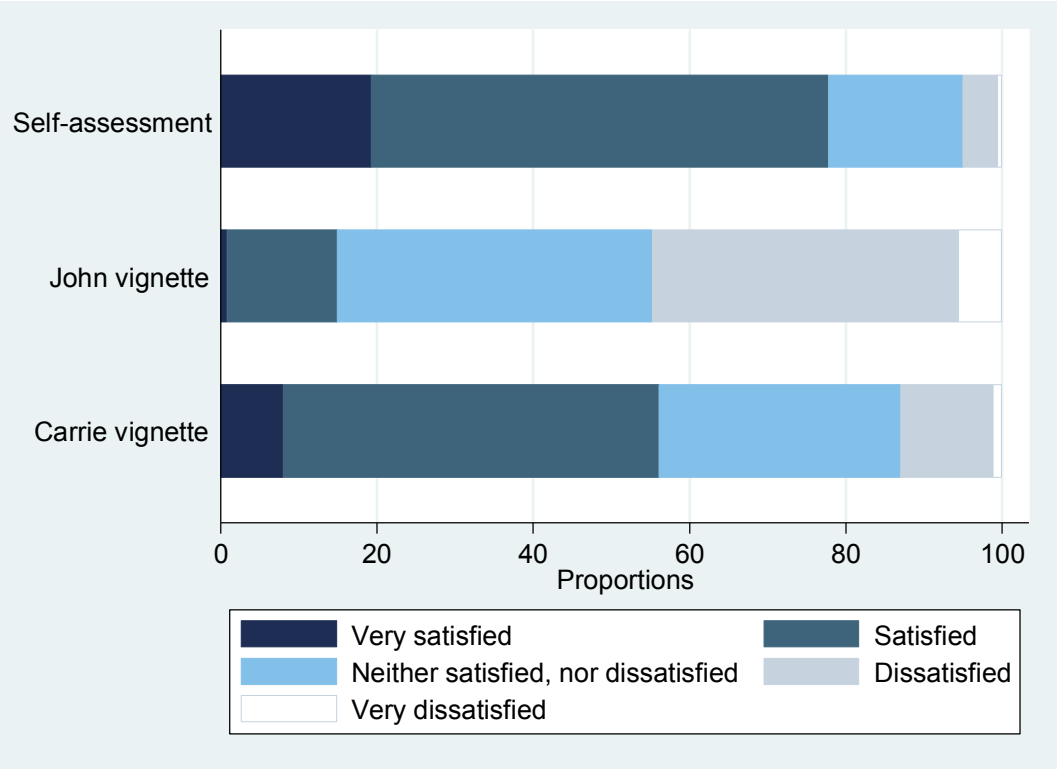
Figures and Tables

Figure 1. Self-assessment and anchoring vignette questions for life satisfaction.



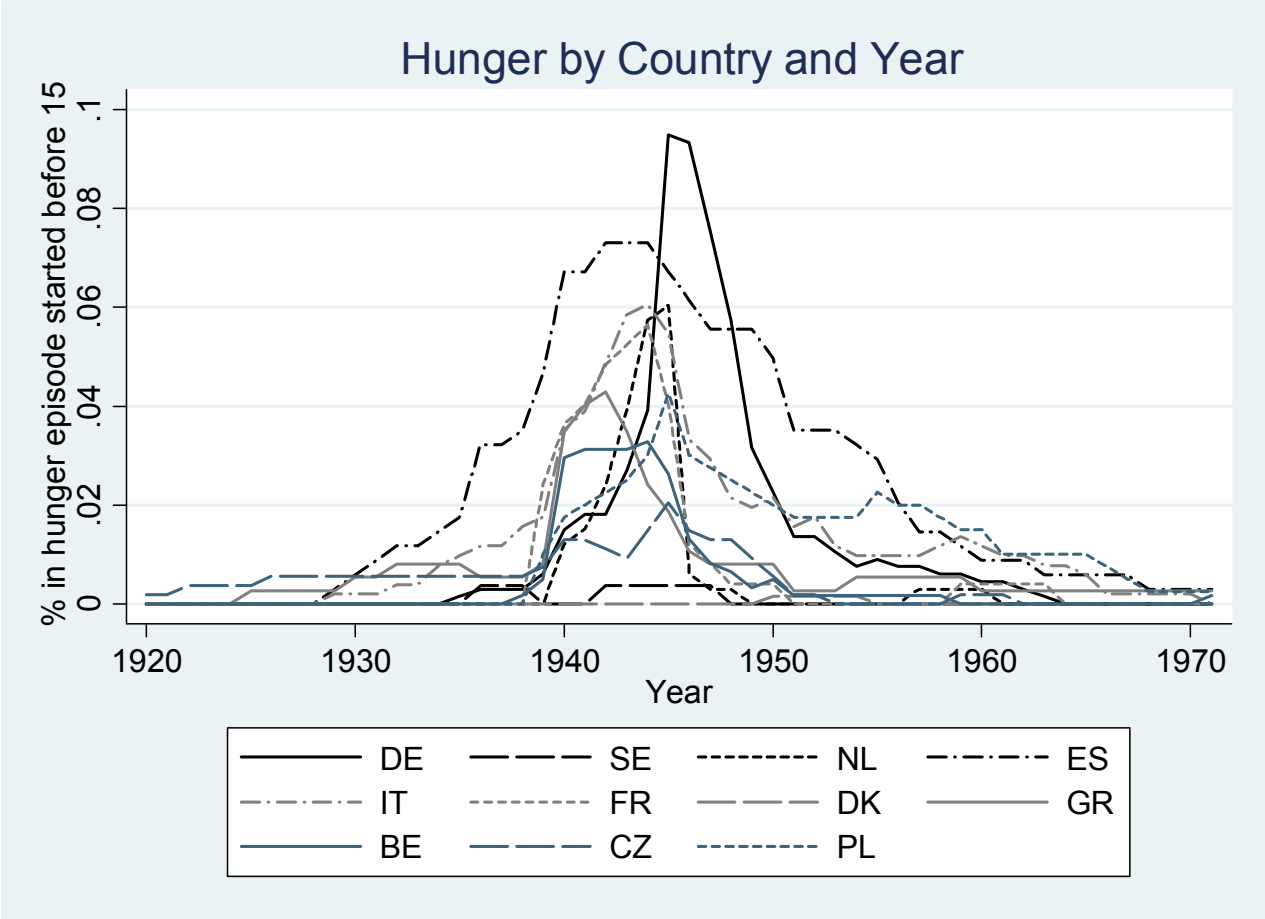
Notes: Monetary values were PPP-adjusted across countries. *Source:* SHARE wave 2 questionnaire.

Figure 2. Self-assessment and vignette evaluations for life satisfaction



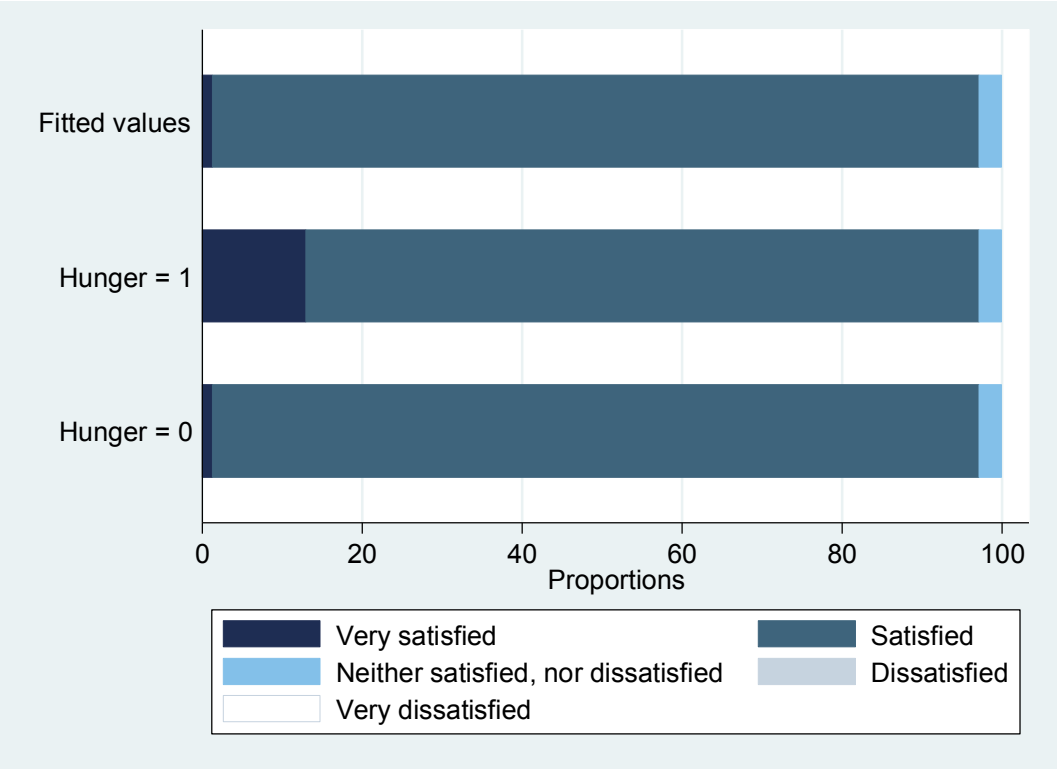
Source: SHARE.

Figure 3. Childhood hunger episodes, by country and year.



Notes: only people in a hunger episode started before age 15 are considered. *Source:* SHARE

Figure 4. Model predictions and counterfactual simulations.



Notes: the upper bar reports the distribution of life satisfaction predicted by the Hopit model. The middle (lower) bar reports the counterfactual distributions of life satisfaction that would hold if all respondents were given the reporting scale they would have experienced had they (had they not) been exposed to hunger episodes in childhood, leaving anything else unchanged. *Source:* SHARE

Table 1. Descriptive statistics.

	Mean	Std. Dev.	Min	Max
Self-assessed life satisfaction	3.915	0.766	1	5
John's vignette	2.655	0.819	1	5
Carrie's vignette	3.500	0.846	1	5
Female	0.545	0.498	0	1
Born before 1930	0.099	0.299	0	1
Born 1930-1934	0.105	0.307	0	1
Born 1935-1939	0.141	0.348	0	1
Born 1940-1944	0.211	0.408	0	1
Born 1945-1949	0.182	0.386	0	1
Born after 1949	0.262	0.440	0	1
BE	0.123	0.329	0	1
CZ	0.109	0.312	0	1
DE	0.134	0.341	0	1
DK	0.135	0.342	0	1
ES	0.069	0.254	0	1
FR	0.050	0.218	0	1
GR	0.075	0.264	0	1
IT	0.103	0.305	0	1
NL	0.067	0.250	0	1
PL	0.081	0.272	0	1
SE	0.053	0.225	0	1
Hunger	0.057	0.231	0	1
Childhood SES	0.000	1.000	-2.022	5.889
Dispossession	0.027	0.163	0	1
Lived in orphanage	0.014	0.116	0	1
Lived with foster parents	0.011	0.104	0	1
Relocation for war	0.034	0.181	0	1
War exposure	0.452	0.498	0	1
Rural area	0.437	0.496	0	1
Troubled parents	0.666	0.472	0	1
Mother absent	0.033	0.178	0	1
Father absent	0.093	0.290	0	1
Siblings	0.831	0.375	0	1
In a couple	0.780	0.414	0	1
Primary education	0.269	0.443	0	1
Secondary education	0.490	0.500	0	1
Higher education	0.241	0.428	0	1
With ADL limitations	0.082	0.275	0	1
With IADL limitations	0.129	0.336	0	1
Retired	0.510	0.500	0	1
Employed	0.288	0.453	0	1
Wealth	6.335	3.153	-14.195	15.694
Income	5.616	2.345	0.775	13.273

Notes: the table reports descriptive statistics for outcome variables, childhood conditions and adulthood conditions in the upper, middle and lower panel, respectively.

Table 2. Childhood hunger and other childhood conditions.

Female	-0.00253	Rural area	-0.0146***
	(0.00301)		(0.00359)
Childhood SES	-0.00740***	Dispossession	0.115***
	(0.00211)		(0.0311)
Troubled Parents	0.0112***	Lived in orphanage	0.0730**
	(0.00314)		(0.0365)
Mother Absent	-0.00357	Lived with foster parents	-0.0109
	(0.00706)		(0.00746)
Father Absent	0.0396***	Relocation for war	0.0397**
	(0.0118)		(0.0155)
Siblings	0.000124	War exposure	0.0252***
	(0.00424)		(0.00908)
Country Dummies	Yes		
Cohort Dummies	Yes		
Observations	4,950		

Notes: the table reports marginal effects from a probit regression. Dependent variable: suffering of hunger in childhood. One, two and three stars for statistical significance of the underlying coefficient at ten, five and one percent levels of confidence.

Table 3. Hunger effects on self-reported wellbeing.

	(1)	(2)	(3)	(4)	(5)	(6)
	Baseline	Hopit	Cut-off 1	Cut-off 2	Cut-off 3	Cut-off 4
Hunger	-0.281***	-0.353***	0.0545	0.0400	-0.0948	-0.195***
	(0.0781)	(0.0929)	(0.109)	(0.0716)	(0.0591)	(0.0553)
Observations	4,950					
Cut-offs (P-value)	0.001					

Notes: the table reports coefficients related with hunger from an extended Hopit model for life satisfaction. Other childhood covariates included are described in Section 1. Full estimation outcomes are reported in Table A2 in the Appendix. The p-value reported in the bottom line refers to a test for joint significance of the hunger coefficients in the four cut-off equations. One, two and three stars for statistical significance at ten, five and one percent levels of confidence.

Table 4. Hunger effects on self-reported wellbeing - with mediators.

	(1)	(2)	(3)	(4)	(5)	(6)
	Baseline	Hopit	Cut-off 1	Cut-off 2	Cut-off 3	Cut-off 4
Hunger	-0.278***	-0.369***	0.0347	0.0480	-0.0899	-0.206***
	(0.0795)	(0.0960)	(0.116)	(0.0719)	(0.0592)	(0.0550)
Observations	4,950					
Cut-offs (P-value)	0.000					

Notes: the table reports coefficients related with hunger from an extended Hopit model for life satisfaction. Other childhood covariates included are described in Section 1. The model includes also the set of adulthood mediators described in Section 1. Full estimation outcomes are reported in Table A3 in the Appendix. The p-value reported in the bottom line refers to a test for joint significance of the hunger coefficients in the four cut-off equations. One, two and three stars for statistical significance at ten, five and one percent levels of confidence.

Table 5. Dynamic selection across SHARE waves, vignette evaluation and life satisfaction

	(1)	(2)
John's vignette	0.002	0.001
	(0.002)	(0.002)
Carrie's vignette	-0.001	-0.001
	(0.002)	(0.002)
Life satisfaction		0.001
		(0.002)
Wave 2 covariates	Yes	Yes
Country dummies	Yes	Yes
Cohort dummies	Yes	Yes
Observations	5,459	5,459

Notes: the table reports marginal effects from Probit regressions. Dependent variable: deceased in SHARE wave 3. The regression controls for country and cohort fixed effects, gender, education, marital status, employment status, income, wealth and health, all measured at wave 2. One, two and three stars for statistical significance at ten, five and one percent levels of confidence.

Table 6. Hunger effects on self-reported wellbeing - excluding older cohorts.

	(1)	(2)	(3)	(4)	(5)	(6)
	Baseline	Hopit	Cut-off 1	Cut-off 2	Cut-off 3	Cut-off 4
1930-1956 cohorts						
Hunger	-0.272***	-0.356***	0.126	0.0180	-0.123*	-0.194***
	(0.0851)	(0.102)	(0.118)	(0.0773)	(0.0654)	(0.0594)
Observations	4,460					
Cut-offs (P-value)	0.001					
1935-1956 cohorts						
Hunger	-0.309***	-0.416***	0.103	0.0307	-0.144*	-0.188***
	-0.103	(0.123)	(0.144)	(0.0935)	(0.0805)	(0.0709)
Observations	3,939					
Cut-offs (P-value)	0.015					

Notes: the table reports coefficients related with hunger from an extended Hopit model for life satisfaction. Other childhood covariates included are described in Section 1. Full estimation outcomes are available from the author. The p-value reported in the bottom line of each panel refers to a test for joint significance of the hunger coefficients in the four cut-off equations. One, two and three stars for statistical significance at ten, five and one percent levels of confidence.

Table 7. Heterogeneous hunger effects on self-reported wellbeing across genders.

	(1)	(2)	(3)	(4)	(5)	(6)
	Baseline	Hopit	Cut-off 1	Cut-off 2	Cut-off 3	Cut-off 4
Males						
Hunger	-0.199*	-0.228*	0.0954	0.0999	-0.102	-0.252***
	(0.112)	(0.136)	(0.152)	(0.0957)	(0.0854)	(0.0818)
Observations	2,252					
Cut-offs (P-value)	0.003					
Females						
Hunger	-0.388***	-0.512***	0.00362	-0.00933	-0.0841	-0.165**
	(0.110)	(0.130)	(0.164)	(0.109)	(0.0828)	(0.0774)
Observations	2,698					
Cut-offs (P-value)	0.092					

Notes: the table reports coefficients related with hunger from an extended Hopit model for life satisfaction. Estimation is carried out separately for males and females. Other childhood covariates included are described in Section 1. Full estimation outcomes are available from the author. The p-value reported in the bottom line of each panel refers to a test for joint significance of the hunger coefficients in the four cut-off equations. One, two and three stars for statistical significance at ten, five and one percent levels of confidence.

Table 8. Heterogeneous hunger effects on self-reported wellbeing for individuals with siblings and single children.

	(1)	(2)	(3)	(4)	(5)	(6)
	Baseline	Hopit	Cut-off 1	Cut-off 2	Cut-off 3	Cut-off 4
With siblings						
Hunger	-0.229***	-0.297***	0.0830	0.0416	-0.0986	-0.223***
	(0.0864)	(0.103)	(0.117)	(0.0766)	(0.0655)	(0.0617)
Observations	4,115					
Cut-offs (P-value)	0.001					
No siblings						
Hunger	-0.473**	-0.573**	0.142	-0.126	-0.0513	-0.0209
	(0.187)	(0.223)	(0.271)	(0.166)	(0.137)	(0.131)
Observations	835					
Cut-offs (P-value)	0.876					

Notes: the table reports coefficients related with hunger from an extended Hopit model for life satisfaction. Estimation is carried out separately for individuals with siblings and single children. Other childhood covariates included are described in Section 1. Full estimation outcomes are available from the author. The p-value reported in the bottom line of each panel refers to a test for joint significance of the hunger coefficients in the four cut-off equations. One, two and three stars for statistical significance at ten, five and one percent levels of confidence.

Table 9. Heterogeneous hunger effects on self-reported wellbeing for individuals coming from high and low socio-economic status in childhood.

	(1)	(2)	(3)	(4)	(5)	(6)
	Baseline	Hopit	Cut-off 1	Cut-off 2	Cut-off 3	Cut-off 4
High SES						
Hunger	-0.0684	-0.0315	0.0466	0.139	0.0399	-0.255**
	(0.144)	(0.186)	(0.202)	(0.121)	(0.104)	(0.102)
Observations	2,473					
Cut-offs (P-value)	0.013					
Low SES						
Hunger	-0.346***	-0.454***	0.126	-0.0574	-0.149**	-0.138**
	(0.0939)	(0.111)	(0.131)	(0.0858)	(0.0721)	(0.0678)
Observations	2,477					
Cut-offs (P-value)	0.016					

Notes: the table reports coefficients related with hunger from an extended Hopit model for life satisfaction. Estimation is carried out separately for individuals above and below the median level of socio-economic status. Other childhood covariates included are described in Section 1. Full estimation outcomes are available from the author. The p-value reported in the bottom line of each panel refers to a test for joint significance of the hunger coefficients in the four cut-off equations. One, two and three stars for statistical significance at ten, five and one percent levels of confidence.

Appendix

Table A1. Attrition between the full COMPARE sample (wave 2) and the subsample of survivors in SHARELIFE (wave 3).

Country	Sample	N. obs	Attrition rate (%)	% female	Age	Log (income)	% higher education	% with ADL limitation	% in a couple	% employed	% retired	Self-assessed life satisf.	John's vignette	Carrie's vignette
Germany	Wave2	1,103	38.1	52.7	64.86	10.07	29.9	6.6	80.3	27.3	54.0	3.9	2.9	3.4
	Wave3	683		52.0	64.55	10.13	33.4	5.7	81.0	27.8	52.7	4.0	2.8	3.4
Sweden	Wave2	440	40.2	54.1	65.54	10.24	33.4	8.2	78.4	37.3	58.4	4.2	2.4	3.4
	Wave3	263		52.5	65.13	10.31	33.5	6.1	79.1	38.8	57.4	4.3	2.3	3.4
Netherlands	Wave2	482	33.0	51.7	61.63	10.36	27.8	3.3	84.9	40.7	33.8	4.1	2.7	3.3
	Wave3	323		52.3	62.22	10.34	28.2	2.8	83.6	36.2	36.8	4.1	2.7	3.3
Spain	Wave2	478	28.9	52.9	64.18	9.41	13.4	8.8	81.4	26.2	34.3	3.8	2.5	3.4
	Wave3	340		55.3	64.35	9.35	12.4	8.2	80.9	25.3	33.2	3.8	2.4	3.5
Italy	Wave2	666	24.2	53.6	64.98	9.74	9.2	8.4	83.6	18.2	53.6	3.6	2.4	3.3
	Wave3	505		53.3	64.84	9.70	8.3	8.9	85.7	17.6	56.4	3.6	2.4	3.3
France	Wave2	356	29.8	55.1	64.28	10.26	25.3	9.8	70.8	27.5	56.7	3.8	2.4	3.2
	Wave3	250		54.8	64.04	10.34	26.8	8.4	69.2	30.8	56.8	3.8	2.4	3.2
Denmark	Wave2	926	27.5	54.3	64.35	10.20	39.5	5.9	81.9	43.2	47.8	4.3	3.0	3.8
	Wave3	671		54.7	64.13	10.25	41.7	5.2	82.0	45.0	47.1	4.4	3.0	3.8
Greece	Wave2	498	25.3	52.2	64.63	9.51	21.1	5.8	71.3	33.3	39.8	3.6	2.7	3.3
	Wave3	372		54.8	64.53	9.45	19.4	4.6	69.1	31.7	39.8	3.6	2.7	3.3
Belgium	Wave2	812	22.5	53.6	65.51	10.01	23.4	10.3	75.0	21.1	51.6	3.9	2.5	3.6
	Wave3	629		54.1	65.87	10.03	22.4	11.3	75.2	19.7	52.3	4.0	2.5	3.7
Czechia	Wave2	850	37.4	58.5	64.31	9.48	11.2	7.5	70.1	27.2	67.3	3.7	2.8	3.7
	Wave3	532		58.6	64.15	9.51	12.8	6.6	70.5	26.1	68.4	3.7	2.8	3.6
Poland	Wave2	527	27.5	56.4	62.93	9.06	17.6	26.2	75.9	18.8	55.4	3.7	2.6	3.5
	Wave3	382		57.3	61.50	9.06	19.9	23.8	76.7	21.7	51.8	3.7	2.7	3.6
Total	Wave2	7,138	30.7	54.2	64.40	9.86	23.5	8.8	77.8	29.0	51.3	3.9	2.7	3.5
	Wave3	4,950		54.5	64.25	9.87	24.1	8.2	78.0	28.8	51.0	3.9	2.7	3.5

Notes: the table reports attrition rate and mean values of several variables measured at baseline (wave 2), for the full wave 2 COMPARE sample and for the selected wave 3 sample we consider in our analysis, for each country and for the pooled sample.

Table A2. Childhood conditions and self-reported wellbeing.

	(1) Baseline	(2) Hopit	(3) Cut 1	(4) Cut 2	(5) Cut 3	(6) Cut 4
Hunger	-0.281***	-0.353***	0.0545	0.0400	-0.0948	-0.195***
	(0.0781)	(0.0929)	(0.109)	(0.0716)	(0.0591)	(0.0553)
Childhood SES	0.104***	0.163***	0.0310	-0.00607	0.00787	0.0317**
	(0.0201)	(0.0253)	(0.0334)	(0.0212)	(0.0153)	(0.0129)
Dispossession	0.0178	0.0368	-0.0917	-0.0296	0.0910	0.0810
	(0.109)	(0.135)	(0.182)	(0.118)	(0.0767)	(0.0698)
Lived in orphanage	0.0606	-0.0527	-0.384	0.130	0.0749	-0.0720
	(0.152)	(0.189)	(0.313)	(0.168)	(0.111)	(0.0971)
Lived with foster parents	-0.0409	0.0397	0.104	0.0316	-0.131	0.0713
	(0.174)	(0.227)	(0.296)	(0.162)	(0.142)	(0.111)
Relocation for war	0.000518	0.164	0.172	0.0730	-0.131*	0.101
	(0.101)	(0.132)	(0.142)	(0.0914)	(0.0794)	(0.0658)
War exposure	-0.0278	-0.0357	0.208*	-0.112*	-0.0191	-0.0164
	(0.0636)	(0.0789)	(0.113)	(0.0674)	(0.0474)	(0.0406)
Rural	-0.0646*	-0.0116	-0.0986*	0.117***	-0.0349	0.0143
	(0.0367)	(0.0453)	(0.0599)	(0.0371)	(0.0274)	(0.0236)
Troubled parents	-0.0569	-0.0995**	-0.0479	0.00474	0.0181	-0.0352
	(0.0378)	(0.0469)	(0.0589)	(0.0372)	(0.0282)	(0.0249)
Mother absent	0.0802	0.0865	-0.165	0.150	-0.0920	0.0109
	(0.104)	(0.127)	(0.169)	(0.0945)	(0.0809)	(0.0678)
Father absent	-0.0383	-0.0459	0.00322	-0.0462	0.111**	-0.0678
	(0.0649)	(0.0799)	(0.107)	(0.0664)	(0.0464)	(0.0424)
Siblings	0.150***	0.149**	-0.143*	0.0280	0.0489	0.0415
	(0.0479)	(0.0586)	(0.0747)	(0.0465)	(0.0355)	(0.0309)
Female	-0.0862**	-0.147***	0.0653	-0.0510	-0.0133	-0.0530**
	(0.0343)	(0.0426)	(0.0547)	(0.0340)	(0.0256)	(0.0221)
Born before 1930	-0.206***	-0.104	0.207*	-0.0701	0.0656	-0.0740
	(0.0719)	(0.0881)	(0.122)	(0.0747)	(0.0531)	(0.0479)
Born 1930-1934	-0.0919	0.0961	0.206	-0.0747	0.0965*	0.0307
	(0.0766)	(0.0956)	(0.130)	(0.0790)	(0.0570)	(0.0497)
Born 1935-1939	0.113	0.261***	0.280**	-0.112	0.0414	0.0138
	(0.0720)	(0.0911)	(0.124)	(0.0766)	(0.0545)	(0.0463)
Born 1945-1949	0.0396	0.0955	-0.0840	0.0687	0.0246	0.00338
	(0.0673)	(0.0845)	(0.123)	(0.0720)	(0.0514)	(0.0425)
Born after 1949	-0.172***	-0.205***	0.0784	-0.0761	0.00965	-0.00847
	(0.0519)	(0.0636)	(0.0892)	(0.0541)	(0.0388)	(0.0323)
SE	0.598***	0.635***	0.546***	0.0243	-0.113	-0.311***
	(0.102)	(0.135)	(0.174)	(0.0997)	(0.0763)	(0.0692)
NL	0.291***	0.555***	0.652***	-0.312***	0.0147	0.0413
	(0.0838)	(0.120)	(0.140)	(0.0940)	(0.0604)	(0.0536)
ES	-0.179**	-0.213**	0.760***	-0.0893	-0.467***	-0.161***
	(0.0827)	(0.103)	(0.137)	(0.0838)	(0.0689)	(0.0530)
IT	-0.479***	-0.321***	0.886***	-0.221***	-0.200***	-0.0286
	(0.0728)	(0.0914)	(0.125)	(0.0771)	(0.0543)	(0.0496)
FR	-0.197**	0.151	0.639***	-0.0350	0.0101	-0.0745
	(0.0898)	(0.117)	(0.148)	(0.0882)	(0.0637)	(0.0644)
DK	0.572***	0.132	-0.324*	0.0188	-0.0893	-0.161***
	(0.0808)	(0.103)	(0.193)	(0.108)	(0.0609)	(0.0490)
GR	-0.479***	-0.529***	0.773***	-0.264***	-0.0104	-0.608***
	(0.0797)	(0.0971)	(0.137)	(0.0879)	(0.0558)	(0.0603)
BE	0.0536	-0.218**	0.442***	-0.0860	-0.284***	-0.286***
	(0.0689)	(0.0870)	(0.129)	(0.0783)	(0.0544)	(0.0440)
CZ	-0.455***	-0.631***	0.0269	-0.0796	0.0464	-0.131***
	(0.0713)	(0.0894)	(0.149)	(0.0899)	(0.0502)	(0.0450)
PL	-0.366***	-0.487***	0.614***	-0.229***	-0.204***	-0.167***
	(0.0787)	(0.0964)	(0.139)	(0.0864)	(0.0595)	(0.0506)
Constant			-4.184***	0.587***	0.212***	0.786***
			(0.191)	(0.0945)	(0.0649)	(0.0547)
Baseline model - Cut-offs			-4.181***	0.590***	0.204***	0.784***
			(0.189)	(0.0934)	(0.0642)	(0.0542)
θ_1	-1.988***	-2.070***				
	(0.0897)	(0.109)				
θ_2	-0.793***	-0.884***				
	(0.0855)	(0.106)				
σ_v	0.0194	0.013				
	(0.0195)	(0.0198)				
σ_η	0.454***	0.406***				
	(0.0183)	(0.0182)				
Observations	4,950					

Notes: the table reports coefficients from an extended Hopit model for life satisfaction. Childhood covariates included are described in Section 1. One, two and three stars for statistical significance at ten, five and one percent levels of confidence.

Table A3. Estimated country dummies and vignette levels in the pooled sample and separating by group of countries identified in accordance with the Inglehart-Wezel values map.

	(1) Full sample	(4) Protestant countries	(3) Catholic countries	(2) Ex-communist countries
DK	0.101	0.170		
	(0.104)	(0.129)		
NL	0.539***	0.557***		
	(0.120)	(0.125)		
SE	0.607***	0.681***		
	(0.136)	(0.156)		
BE	-0.224***		-0.213**	
	(0.0866)		(0.0883)	
ES	-0.211**		-0.193*	
	(0.102)		(0.105)	
FR	0.150		0.133	
	(0.117)		(0.117)	
IT	-0.317***		-0.306***	
	(0.0910)		(0.0937)	
CZ	-0.632***			-0.655***
	(0.0891)			(0.0945)
PL	-0.487***			-0.477***
	(0.0961)			(0.104)
θ_1	-2.078***	-1.758***	-1.966***	-1.909***
	(0.112)	(0.179)	(0.145)	(0.169)
θ_2	-0.862***	-0.669***	-0.793***	-0.794***
	(0.109)	(0.175)	(0.141)	(0.164)
Observations	4,577	1,928	2,375	1,602

Notes: the table reports the coefficient on country dummies and on the rating levels of each vignette question from an extended Hopit model for life satisfaction. Estimation is carried out on the full set of countries and on three separate groups defined in accordance with the Inglehart-Wezel values map. Since it is an orthodox country, Greece is excluded from this analysis. Protestant countries are Denmark, Germany, the Netherlands and Sweden. Catholic countries include Belgium, Italy, France and Spain. Czech Republic and Poland are included amongst ex-communist countries. All samples include Germany as baseline country. Full estimation outcomes are available from the author. Other childhood covariates included are described in Section 1. One, two and three stars for statistical significance at ten, five and one percent levels of confidence.

Table A4. Childhood conditions, adulthood mediators and self-reported wellbeing.

	(1) Baseline	(2) Hopit	(3) Cut-off 1	(4) Cut-off 2	(5) Cut-off 3	(6) Cut-off 4
Hunger	-0.278***	-0.369***	0.0347	0.0480	-0.0899	-0.206***
	(0.0795)	(0.0960)	(0.116)	(0.0719)	(0.0592)	(0.0550)
Childhood SES	0.0694***	0.118***	0.0293	-0.0103	0.0123	0.0172
	(0.0215)	(0.0274)	(0.0366)	(0.0216)	(0.0160)	(0.0134)
Dispossession	0.0211	0.0558	-0.0824	-0.0404	0.0942	0.0941
	(0.110)	(0.140)	(0.194)	(0.118)	(0.0767)	(0.0694)
Lived in orphanage	0.0982	-0.0250	-0.476	0.160	0.0925	-0.0869
	(0.155)	(0.195)	(0.342)	(0.170)	(0.110)	(0.0961)
Lived with foster parents	-0.0478	0.0465	0.0954	0.0368	-0.130	0.0724
	(0.177)	(0.236)	(0.315)	(0.162)	(0.143)	(0.110)
Relocation for war	-0.0562	0.0859	0.168	0.0800	-0.134*	0.0680
	(0.103)	(0.136)	(0.151)	(0.0910)	(0.0792)	(0.0654)
War exposure	-0.0477	-0.0610	0.213*	-0.113*	-0.0126	-0.0175
	(0.0648)	(0.0817)	(0.117)	(0.0661)	(0.0472)	(0.0403)
Rural	-0.0579	0.00436	-0.0980	0.114***	-0.0322	0.0186
	(0.0376)	(0.0472)	(0.0644)	(0.0376)	(0.0275)	(0.0236)
Troubled parents	-0.0387	-0.0746	-0.0621	0.00799	0.0196	-0.0251
	(0.0385)	(0.0485)	(0.0625)	(0.0371)	(0.0282)	(0.0247)
Mother absent	0.105	0.124	-0.169	0.155	-0.105	0.0336
	(0.105)	(0.132)	(0.182)	(0.0956)	(0.0810)	(0.0675)
Father absent	-0.0476	-0.0648	-0.00153	-0.0409	0.107**	-0.0774*
	(0.0661)	(0.0824)	(0.114)	(0.0665)	(0.0463)	(0.0421)
Siblings	0.150***	0.144**	-0.156**	0.0302	0.0473	0.0406
	(0.0488)	(0.0607)	(0.0792)	(0.0465)	(0.0354)	(0.0308)
Female	0.0928**	0.0719	0.0437	-0.0360	-0.00580	-0.00411
	(0.0372)	(0.0469)	(0.0630)	(0.0368)	(0.0273)	(0.0232)
In a couple	0.533***	0.575***	0.171	-0.0746	-0.0402	0.0808*
	(0.0644)	(0.0775)	(0.108)	(0.0630)	(0.0451)	(0.0423)
Primary education	-0.144**	-0.251***	-0.0621	-0.0132	0.0370	-0.116***
	(0.0599)	(0.0756)	(0.0983)	(0.0571)	(0.0445)	(0.0381)
Secondary education	-0.123***	-0.152**	-0.0251	-0.0356	0.0638*	-0.0152
	(0.0463)	(0.0606)	(0.0813)	(0.0468)	(0.0348)	(0.0284)
With ADL limitations	-0.584***	-0.462***	0.228**	-0.0625	-0.0352	0.0807
	(0.0704)	(0.0832)	(0.107)	(0.0671)	(0.0509)	(0.0505)
With IADL limitations	-0.443***	-0.480***	0.0448	0.00315	-0.0461	-0.108***
	(0.0581)	(0.0691)	(0.0924)	(0.0563)	(0.0415)	(0.0397)
Retired	0.226***	0.257***	0.0183	-0.0109	0.00720	0.0589*
	(0.0535)	(0.0657)	(0.0858)	(0.0524)	(0.0394)	(0.0348)
Employed	0.354***	0.428***	-0.102	0.0743	0.0135	0.0662*
	(0.0579)	(0.0718)	(0.105)	(0.0619)	(0.0428)	(0.0366)
Wealth	0.0192**	0.0237**	0.000842	-0.00265	0.00444	0.00893*
	(0.00823)	(0.00966)	(0.0126)	(0.00675)	(0.00581)	(0.00522)
Income	0.0208	0.0186	0.0244	-0.00902	-0.0126	-0.000435
	(0.0143)	(0.0171)	(0.0225)	(0.0127)	(0.0101)	(0.00927)
Constant			-3.668***	0.765***	0.285***	0.641***
			(0.313)	(0.155)	(0.109)	(0.0983)
Baseline model - Cut-offs			-2.980***	-1.364***	-0.126	1.794***
			(0.160)	(0.152)	(0.150)	(0.153)
θ_1	-1.226***	-1.214***				
	(0.152)	(0.186)				
θ_2	0.0427	0.0438				
	(0.151)	(0.185)				
σ_v	0.0817***	0.0460**				
	(0.0198)	(0.0203)				
σ_η	0.475***	0.422***				
	(0.0194)	(0.0192)				
Observations	4,950	4,950				

Notes: the table reports coefficients from an extended Hopit model for life satisfaction. Childhood covariates and adulthood mediators included are described in Section 1. Country and cohort effects are included as well. One, two and three stars for statistical significance at ten, five and one percent levels of confidence.

Table A5. Hunger effects on self-reported wellbeing - sensitivities

	(1)	(2)	(3)	(4)	(5)	(6)
	Baseline	Hopit	Cut-off 1	Cut-off 2	Cut-off 3	Cut-off 4
A. Excluding migrants						
Hunger	-0.312***	-0.394***	0.0590	0.0279	-0.0892	-0.208***
	(0.0799)	(0.0946)	(0.111)	(0.0736)	(0.0604)	(0.0568)
Observations	4,873					
Cut-offs (P-value)	0.001					
B. War countries only						
Hunger	-0.295***	-0.351***	0.0403	0.0562	-0.0989*	-0.181***
	(0.0787)	(0.0944)	(0.111)	(0.0725)	(0.0598)	(0.0559)
Observations	4,017					
Cut-offs (P-value)	0.003					
C. 1920-1949 cohorts						
Hunger	-0.258***	-0.315***	0.0452	0.0647	-0.105*	-0.193***
	(0.0798)	(0.0960)	(0.113)	(0.0736)	(0.0606)	(0.0570)
Observations	3,652					
Cut-offs (P-value)	0.001					

Notes: each panel reports coefficients related with hunger from an extended Hopit model for life satisfaction. Panel A excludes people who migrated from country of birth. Panel B excludes people from Sweden and Denmark. Panel C excludes individuals born after 1949. Other childhood covariates included are described in Section 1. Full estimation outcomes are available from the author. The p-value reported in the bottom line of each panel refers to a test for joint significance of the hunger coefficients in the four cut-off equations. One, two and three stars for statistical significance at ten, five and one percent levels of confidence.